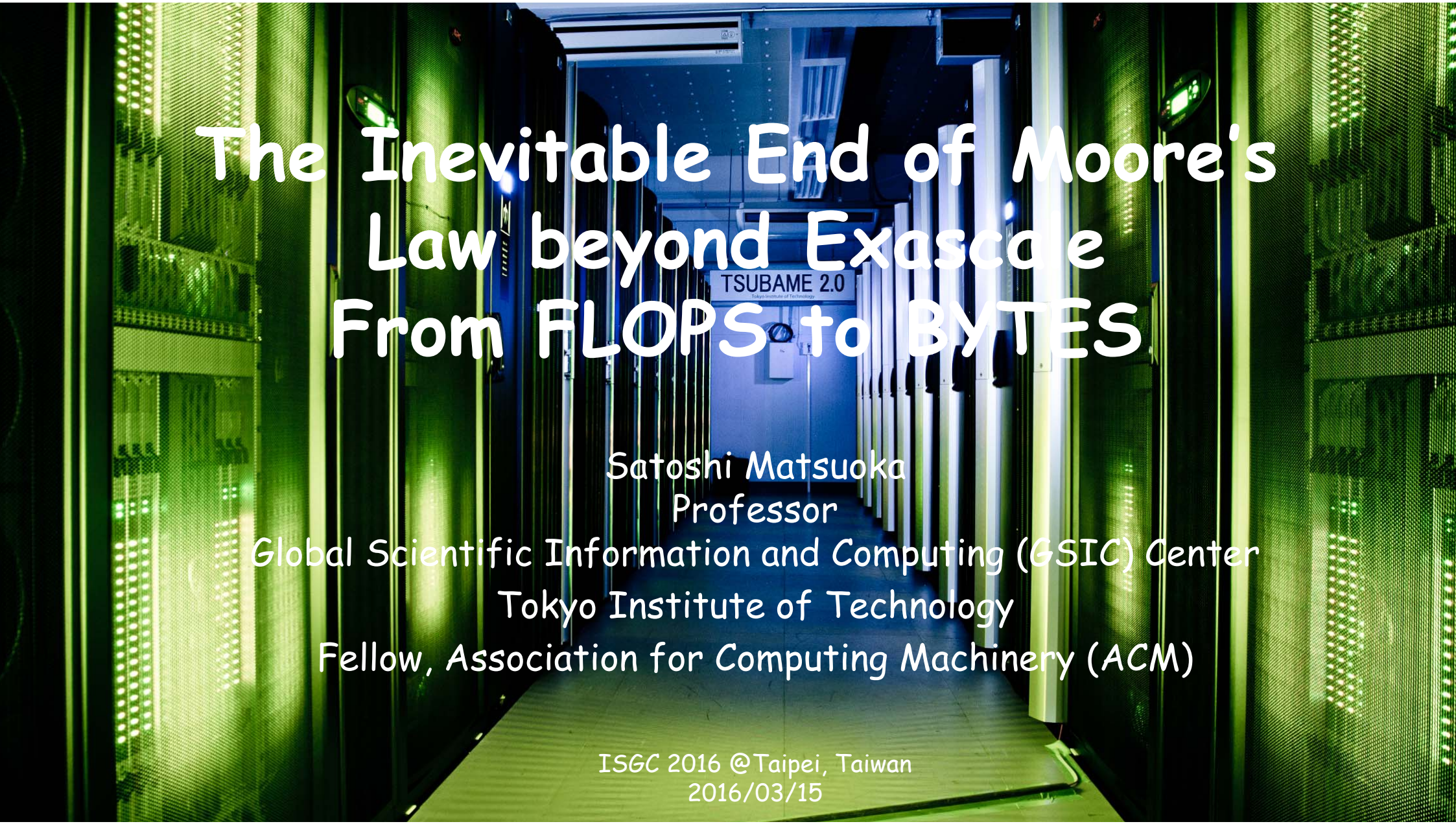


The Inevitable End of Moore's Law beyond Exascale From FLOPS to BYTES

Satoshi Matsuoka
Professor
Global Scientific Information and Computing (GSIC) Center
Tokyo Institute of Technology
Fellow, Association for Computing Machinery (ACM)

ISGC 2016 @Taipei, Taiwan
2016/03/15

Post Moore is the Next "Moonshot" Challenge for the IT and the Society at Large beyond 2020

- Constant transistor power for compute limiting IT performance growth and resulting social innovations
- Need to find the next growth parameters other than **FLOPS=>BYTES**
- Data-capacity and bandwidth (BYTES) increasing device technologies (NVM, **Optics…**) and packaging technologies(3D…)
- BYTES(Bandwidth&Capacity)rich & domain-specifiable architectures
- Programming Paradigm and System Software for bandwidth-rich, asynchronous, data-oriented program acceleration
- Algorithms and Applications that accelerate existing apps (e.g. PDEs) with BYTES, not FLOPS,

POST-MOORE is not just about devices, but a challenge to the whole IT

"From FLOPS to BYTES"

TSUBAME2.0 Nov. 1, 2010
"The Greenest Production Supercomputer in the World"

- GPU-centric (> 4000) high performance & low power
- Small footprint (~200m2 or 2000 sq.ft), low TCO
- High bandwidth memory, optical network, SSD storage…

TSUBAME 2.0
New Development

Chip
(CPU ,GPU)

CPU(Westmere EP)
76.8 GFLOPS
32nm

GPUs(Tesla M2050)
515 GFLOPS
3 GB
40nm

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB
>400GB/s Mem BW
80Gbps NW BW
~1KW max

Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB
>1.6TB/s Mem BW

Rack
(8 Node Chassis)

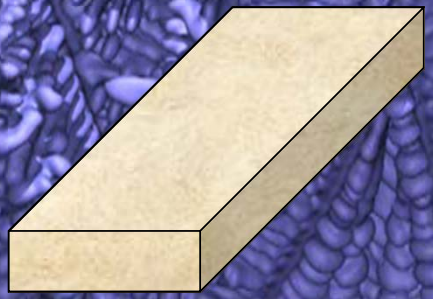
53.6 TFLOPS
1.7 TB/3.2 TB
>12TB/s Mem BW
35KW Max

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes

2.4 PFLOPS
80 TB
4224 GPUs
>600TB/s Mem BW
220Tbps NW
Bisecion BW
1.4MW Max

Integrated by NEC Corporation

ACM Gordon Bell Prize 2011
2.0 Petaflops Dendrite Simulation



Lightweight and durable metal alloy for future automobiles

Special Achievements in Scalability and Time-to-Solution
"Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer"

2 Petaflops (3.4 Petaflops on TSUBAME 2.5)

Comparing K Computer to TSUBAME 2.5

東京工業大学 Tokyo Institute of Technology

GSIC Global Scientific Information and Computing Center

TSUBAME2.0(2010)
→ TSUBAME2.5(2013)
17.1 Petaflops SFP
5.76 Petaflops DFP
$45mil / 6 years (incl. power)

Perf ≒
Cost ≪

独立行政法人理化学研究所 RIKEN
計算科学研究機構
RIKEN Advanced Institute for Computational Science

K Computer (2011)

11.4 Petaflops SFP/DFP
$1400mil 6 years
(incl. power)
x30 TSUBAME2

2016 H2 TSUBAME3.0 Leading Machine Towards Exa & Big Data

1. "Everybody's Supercomputer" - High Performance (15~20 Petaflops, 4~5PB/s Mem, ~Pbit/s NW), innovative high cost/performance packaging & design, in mere 180m^2…
2. "Extreme Green" – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir load leveling & energy recovery
3. "Big Data Convergence" – Extreme high BW &capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, …
4. "Cloud SC" – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds…
5. "Transparency" - full monitoring & user visibility of machine & job state, accountability via reproducibility

2006 TSUBAME1.0
80 Teraflops, #1 Asia #7 World
"Everybody's Supercomputer"

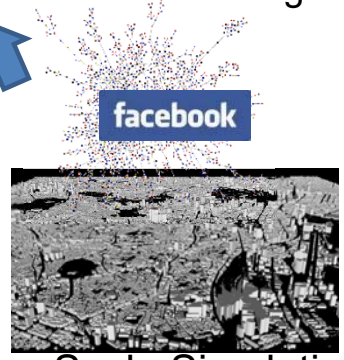
2010 TSUBAME2.0
2.4 Petaflops #4 World
"Greenest Production SC"

2011 ACM Gordon Bell Prize

2013
TSUBAME2.5
upgrade
5.7PF DFP
/17.1PF SFP
20% power
reduction

2013 TSUBAME-KFC
#1 Green 500

2016 TSUBAME3.0
~20PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence



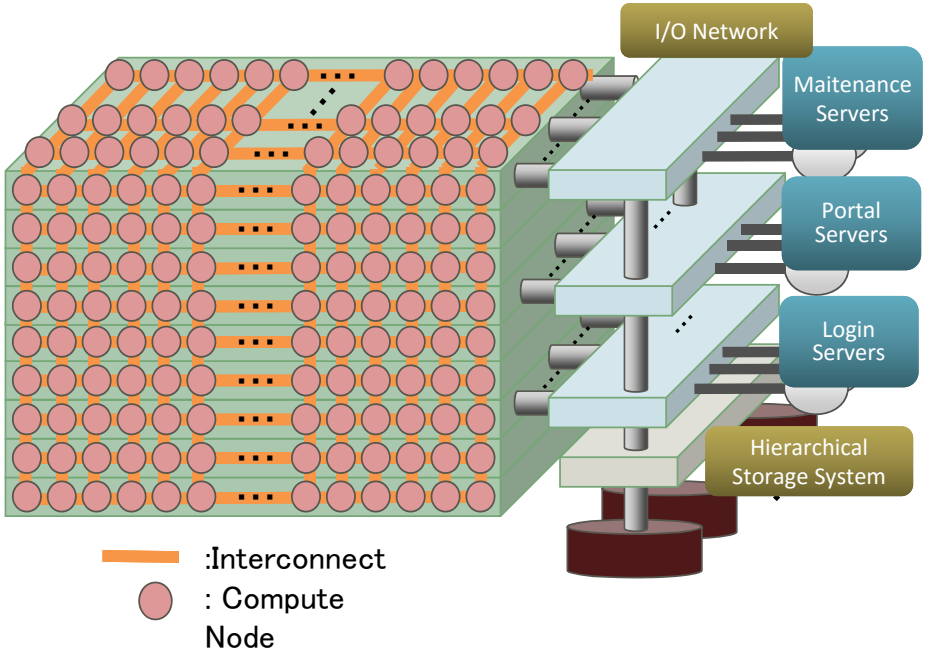
Large Scale Simulation
Big Data Analytics
Industrial Apps

6

Japan Flagship 2020 “Post K” Supercomputer

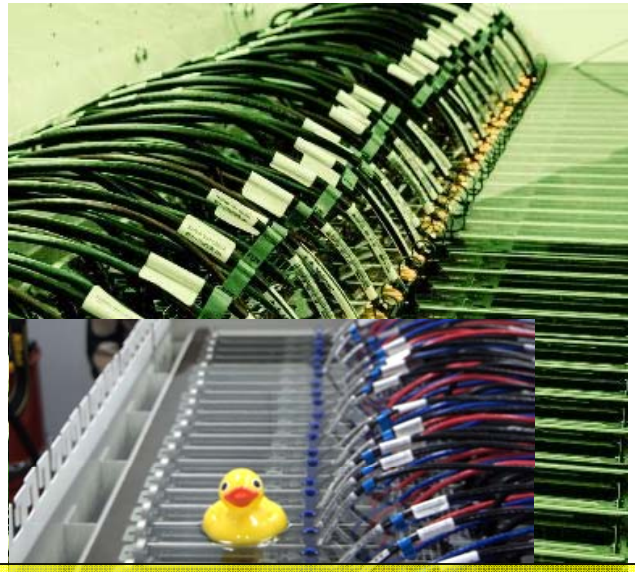
- ✓ CPU
  - A NEW many-core processor (NOT x86)
  - Multi-hundred petaflops peak total
  - Power Knob feature for saving power
- ✓ Memory
  - ✓ 3-D stacked DRAM, Terabyte/s BW
- ✓ Interconnect
  - TOFU3 CPU-integrated 6-D torus network
- I/O acceleration
- 30MW+ Power
- 140 Billion Yen (~1.15 Billion US$)
- Will replace the K computer at the Riken AICS Kobe facility in 2020
- Being designed and will be manufactured by Fujitsu

Prime Minister Abe visiting K Computer 2013



TSUBAME-KFC: TSUBAME3 Prototype [ICPADS2014]

Oil Immersive Cooling＋ Hot Water Cooling + High Density Packaging + Fine-Grained Power Monitoring and Control¥



High Density Oil Immersion
210TFlops (DFP)
630TFlops (SFP)
(=>1.5 PetaFlops upgrade)

High Temperature Cooling
Oil Loop 35~45℃
⇒ Water Loop 25~35℃
(c.f. TSUBAME2: 7~17℃)

Cooling Tower: Water 25~35℃
⇒ To Ambient Air

Experimental Container Facility
20 feet container (16m²)
Fully Unmanned Operation

TSUBAME-KFC #1 Green 500 List (Nov. 2013)

- 1st achievement as Japanese supercomputer
- #1 again in June 2014
- #2 Nov. 2015 w/K80 upgrade

The Green500 List

Listed below are the November 2013 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

TSUBAME-KFC

TSUBAME 2.5

* Performance data obtained from publicly available sources including TOP500

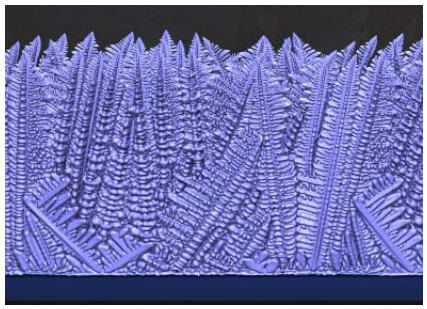
Ranks 1 - 100
Ranks 101 - 200
Ranks 201 - 300
Ranks 301 - 400
Ranks 401 - 500

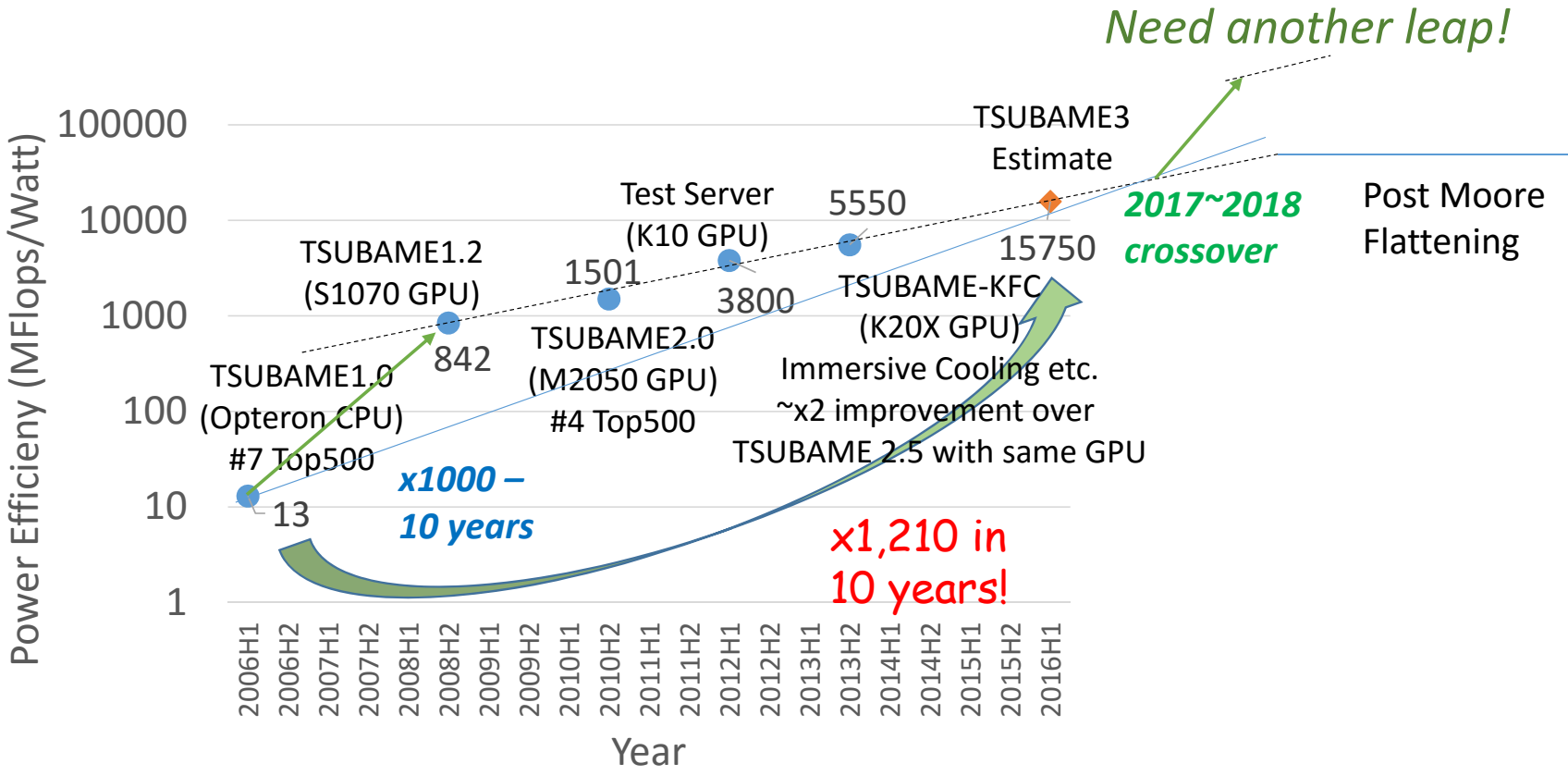
Downloads

Microsoft Excel™ Spreadsheet (.xls)

Comma Separated Values (.csv)

Power Efficiency of GB Dendrite Simulation since 2006
Good news and bad news towards Post-Moore





Measured for the 2011 Gordon Bell Award Dendritic Solidification App
Flop/s/W = Total #Flops / J = energy to solution given same problem

Towards TSUBAME4 and 5: Moore's Law will end in the 2020's

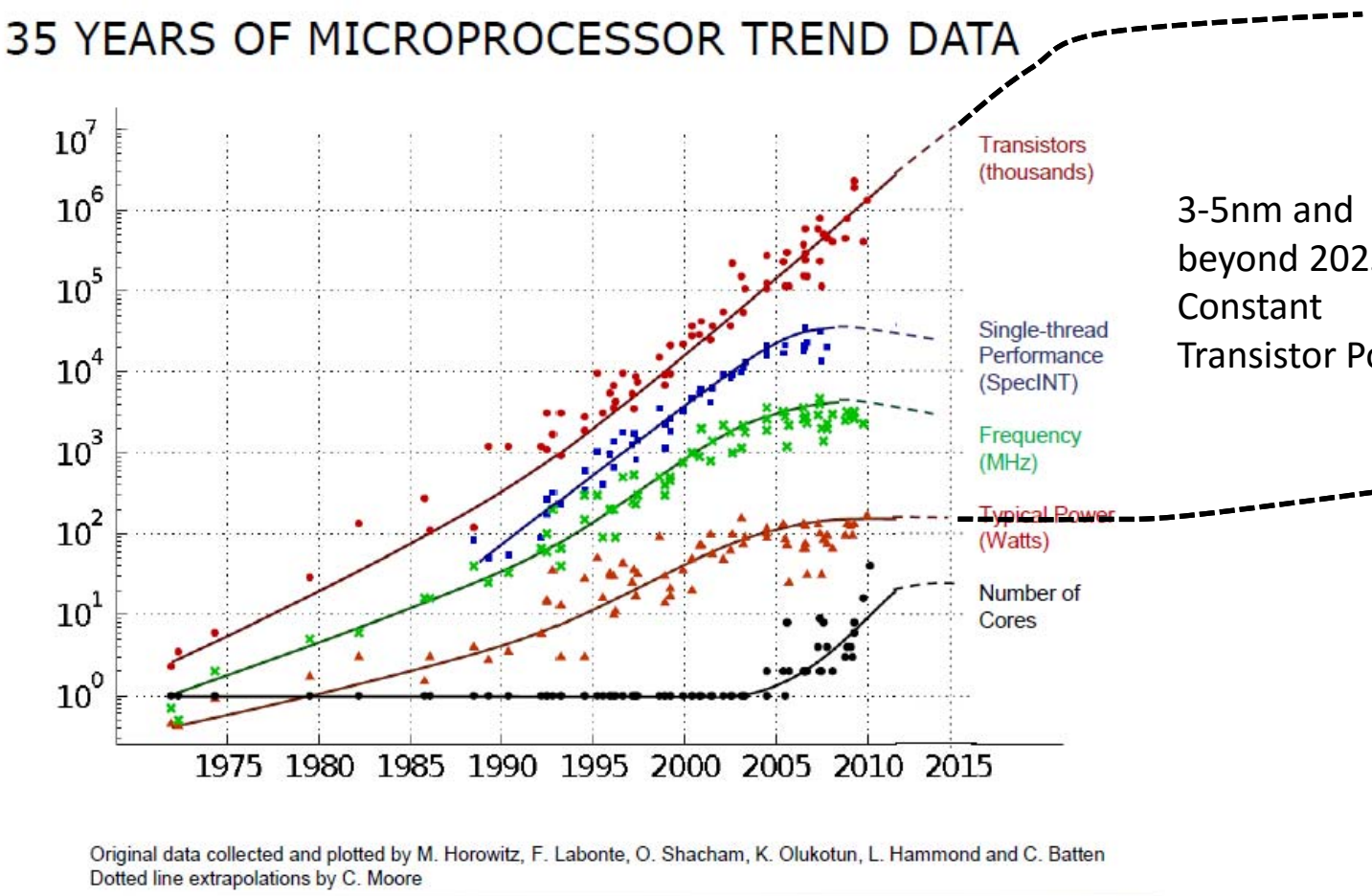
- Much of underlying IT performance growth due to Moore's law
  - "LSI: x2 transistors in 1~1.5 years"
  - Causing qualitative "leaps" in IT and societal innovations
  - The main reason we have supercomputers and Google...
- But this is slowing down & ending, by mid 2020s…!!!
  - End of Lithography shrinks
  - End of Dennard scaling
  - End of Fab Economics

Gordon Moore

The curse of constant transistor power shall soon be upon us

- How do we *sustain* "performance growth" beyond the "end of Moore"?
  - Not just one-time speed bumps
  - Or do we give up and so something else?

The trends will change past End of Moore's Law



- ~2004 Dennard scaling, perf+ = single thread+ = transistor & freq+ = power+
- 2004~2015 feature scaling, perf+ = transistor+ = core#+, constant power
- 2015~2025 all above gets harder
- 2025~ post-Moore, **constant feature&power = flat performance**

End of massive parallelism and many-core speedup eras

The “curse of constant transistor power” when Moore’s Law ends..

- Systems people have been telling the algorithm people that “FLOPS will be free, bandwidth is important, so devise algorithms under that assumption”
- This will certainly be true until exascale in 2020…
- But when Moore’s Law ends in 2025-2030, constant transistor power (esp. for logic) = FLOPS will no longer be free!
- So algorithms that simply increase arithmetic intensity will no longer scale beyond that point
- Have we been telling lies? NO! but is the consequence of device physics trend and the existing system design

But we have new opportunities for post-Moore

- New non-volatile, dense memory devices for massive capacity
- New 3D stacking technologies for massive bandwidth
- Problem specific architectures and reconfigurable archtectures
- Terabit long-range carrier optics into short reach LANs
- New HW computing models e.g. neuromorphic, quantum/Izzling, automata, …

Bold and Controversial Technology Prediction:
Performance growth via *data-centric computing*

- Because data-related parameters (e.g. capacity and bandwidth) *will still continue to grow beyond 2020s (but not compute)*
- Can grow transistor# for compute, but CANNOT use them AT THE SAME TIME(Dark Silicon) => **multiple computing units specialized to type of data**
- **Continued capacity growth** with 3D stacking (esp. direct silicon layering) and low power NVM (e.g. ReRAM)
- Data movement energy initially a problem but will be **capped constant** by dense 3D design and advanced optics from long-haul carrier technologies
- Almost back to the old "vector" days(?), but no free lunch – latency still problem, locality still important, *need **general algorithmic acceleration thru data capacity and bandwidth,** not FLOPS*
- *What are the implications to sparse numerical algotithms?*

Already Enterprise is Reducing Machine Requirements Significantly with Data Capacity Acceleration

2015 - Summary Performance Comparison by Storage Class

	Storage Class			IOPS FOB T4Q32	IOPS Steady State PTS-C T2Q16; PTS-E T4Q32; NVDIMM T128Q4			Bandwidth T1Q32; NVDIMM T32Q1		Response Time T1Q1	
	Category	Device Type	Capacity	RND 4KiB 100% W	RND 4KiB 100% W	RND 4KiB 65:35 RW	RND 4KiB 100% R	SEQ 128KiB 100% W	SEQ 128KiB 100% R	RND 4KiB 100% W Ave	RND 4KiB 100% W Max
	HDD & SSHD - PTS-C v1.2										
1	SSHD	7,200 RPM 2.5" SATA Hybrid	500 GB	3,398	77	62	113	79 MB/s	76 MB/s	12.55 mSec	163.10 mSec
2	SAS HDD	15,000 RPM 2.5" SAS HDD	146 GB	652	133	248	514	147 MB/s	174 MB/s	2.24 mSec	24.54 mSec
	CLIENT SSDs - PTS-C v1.2										
3	mSATA	mSATA 1.8" MLC	250 GB	63,127	5,927	13,545	79,423	300 MB/s	526 MB/s	0.13 mSec	11.86 mSec
4	M.2 x4 AHCI	M.2 x4 Gen 3 2280 MLC	128 GB	77,757	6,307	17,041	172,881	623 MB/s	1,895 MB/s	0.16 mSec	54.19 mSec
5	SATA	SATA6Gb/s 2.5" MLC	400 GB	52,723	34,406	50,646	61,178	428 MB/s	475 MB/s	0.082 mSec	3.13 mSec
	ENTERPRISE SSDs - PTS-E v1.1										
6	SATA	SATA 12Gb/s 2.5" eMLC	800 GB	63,185	38,478	46,911	83,648	430 MB/s	507 MB/s	0.05 mSec	0.99 mSec
7	M.2 x4 NVMe	M.2 x4 Gen3 2280 eMLC WCE	256 GB	29,210	12,155	31,695	248,929	1,100 MB/s	2,158 MB/s	0.09 mSec	23.21 mSec
8	SAS	SAS 12Gb/s 2.5" eMLC	400 GB	149,731	97,167	133,738	187,262	651 MB/s	1,051 MB/s	0.072 mSec	4.72 mSec
9	U.2 NVMe	SFF 8639 4 lane 2.5" MLC	1600 GB	470,122	113,260	238,143	585,884	1,286 MB/s	2,232 MB/s	0.032 mSec	1.00 mSec
10	PCIe x8 AHCI	PCIe 8 Lane Edge Card MLC	1400 GB	160,164	83,021	234,205	750,846	614 MB/s	2,903 MB/s	0.014 mSec	0.56 mSec
	NVDIMM-N - PTS-E v1.1										
11	NVDIMM-N x4	Memory Channel DDR4 x4 Modules	32 GB	2,384,692	2,792,176	2,789,921	3,002,085	35,453 MB/s	55,111 MB/s	0.0022 mSec	0.48 mSec

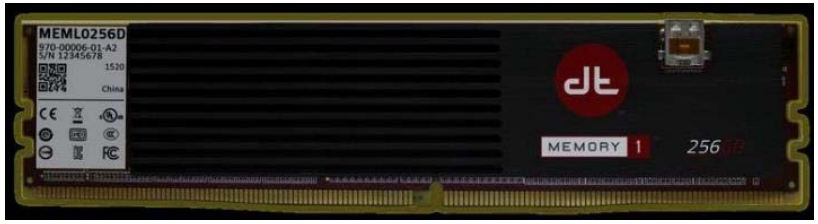
All measurements (rows 1-10) taken by Calypso on the RTP/CTS pursuant to SNIA PTS-E v1.1 / PTS-C v1.2.
IOPS Thread and Queue settings are T2Q16 for HDD/SSHD & Client SSDs, T4Q32 for Enterprise SSDs and T128Q4 for NVDIMM-N.
NVDIMM-N data taken on SuperMicro X10DRI, dual E5 2670V3 8 core cpus, 16GB DRAM using Intel Open Source block IO driver.

SNIA Solid State Storage Initiative — CALYPSO Systems

Source: http://www.storagenewsletter.com/wp-content/uploads/2015/09/sniaSSSINVDIMM-N.jpg

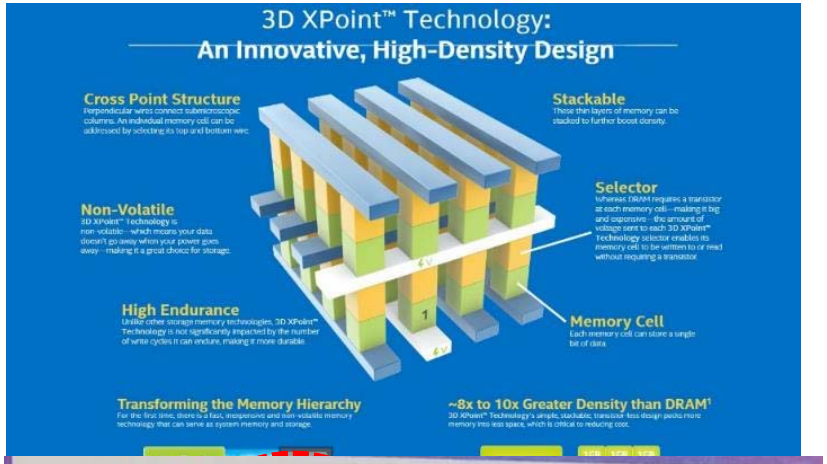
	Current Search Solution	Memory1 Enhanced Solution
Total Index Size	10,000 TB	10,000 TB
Required Servers	100,000	10,000
Memory/Server	128 GB	128 GB + 1 TB Memory1
Index Size/Server	100 GB	1000 GB
Cost/Server	$3,000	$8000
Total Server CAPEX	**$300,000,000**	**$80,000,000**

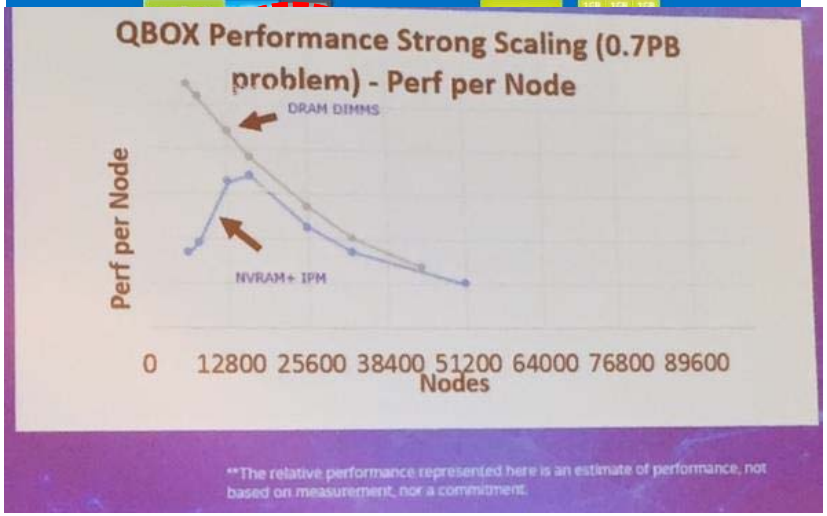
Source: http://www.storagenewsletter.com/wp-content/uploads/2015/09/sniaSSSINVDIMM-N.jp



DIABLO Memory 1 256GB NVDIMM-N

Accelerating “Big Data” Non-Volatile Memory-based Exascale Architectures

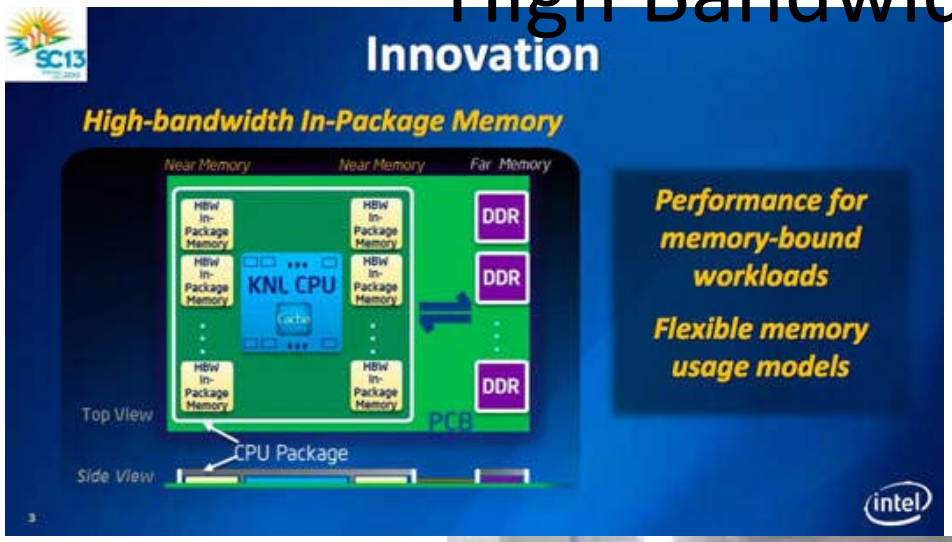




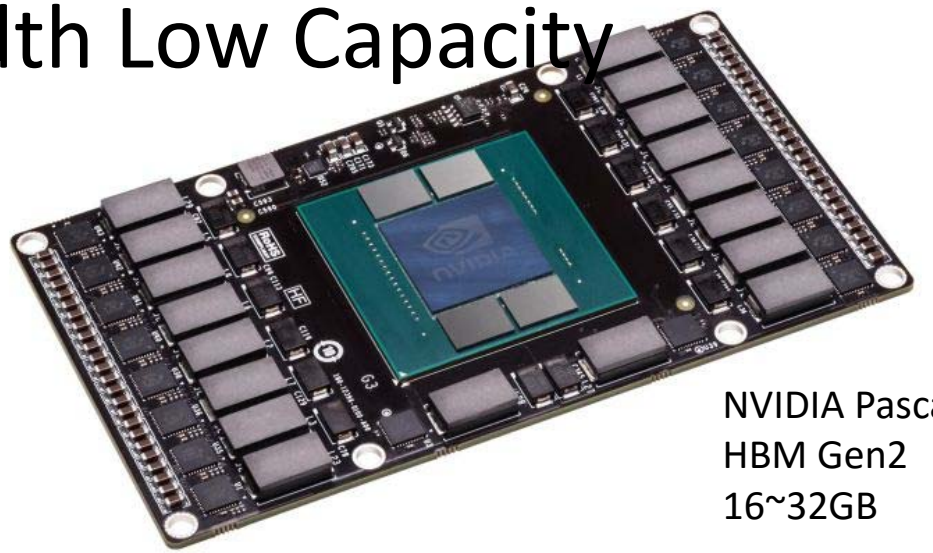
From Al Gara Keynote, IEEE Custer 2015

- Accelerating
  - HPC Apps, Big Data-HPC Apps
- Using
  - Exascale machines w/NVM (Flash, ReRAM, 3D Xpoint, …)
  - BYTES, not FLOPS
- While
  - Reducing Bandwidth, Exploiting Locality
  - Dealing with higher write cost
  - Dealing with low durability
  - Maintaining Programmability
  - Exploiting other system assets such as hybrid electro-optical Exabit interconnect
  - …

Next Gen High BW 3D Stacked Memory
High Bandwidth Low Capacity



Intel KNL + In package memory 16GB



NVIDIA Pascal + HBM Gen2 16~32GB

Fujitsu SPARC fxIX + HMC 32GB

Properties of Post-Moore Architectures: from FLOPS to BYTES

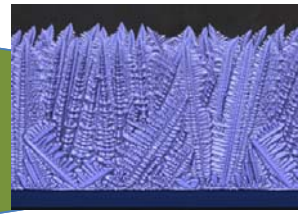
- Constant power for compute => can't increase # transistors (FLOPS) => **BYTES-oriented architecture**
- **BYTES oriented 1: Higher capacity via NVM**
  - Higher density c.f. DRAM, no retention power
  - Can be stacked densely and in many layers in 3D
  - Thinner wafers, inherent 3D structures, TSV for memory...
- **BYTES oriented 2: Higher bandwidth via innovative 3D**
  - 3D&Wafer thinning -> shorter distance
  - Extremely dense TSV with low frequency signalling
  - Other 3D interconnects e.g. inductive/capacitance coupling
  - eventual 3D silicon-on-silicon
- **BYTES oriented 3: Advanced all-optics switched network**
  - Constant power -> one-time E-O O-E conversion, distance oblivious
  - Extremely high bits/W -> carrier-grade technologies into LAN
  - Need to manage circuit switching -> hybrid E-packet/O-circuit network
- **BYTES oriented 4: domain-specific architectural specialization & packaging**
  - Extension of FPGA & SoC
  - Dynamically reconfigurable according to domain needs
  - Low power domain-specific acceleration
  - Including PIM in-memory processing

Other properties such as

- Embedded (tight) packaging
- Fail-in-place architecture
- Immersive cooling
- Extensive power/energy control

Co-Designing Post-Moore HPC System Architecture

FLOPS-Oriented => BYTES-Oriented
Numerical Applications and Algorithms



System SW and Comm MW for Exabit Optical Inteconnect

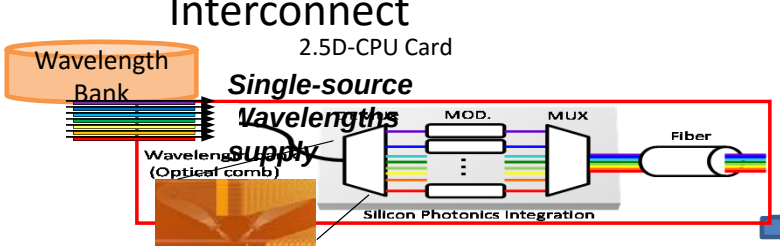
Programming & Perf Modeling of Super Building Block Architecture

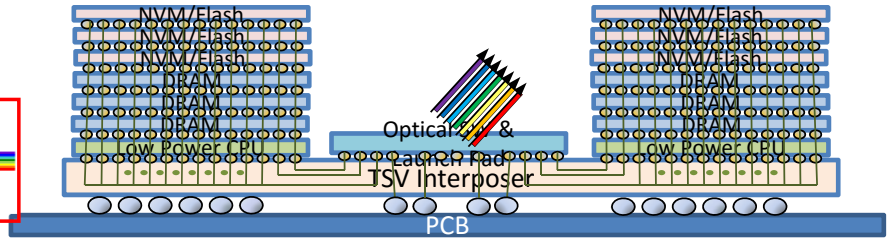
System SW and & Programming of Deep and High Bandwidth Memory Hiearchy

Next Gen Exabit-class Optically Switched Interconnect

Specialized/Integrated/Reconfigurable Super Building Block Architecture

Advanced 3-D stacked Non-Volatile Memory >Tbytes, >10TByte/s



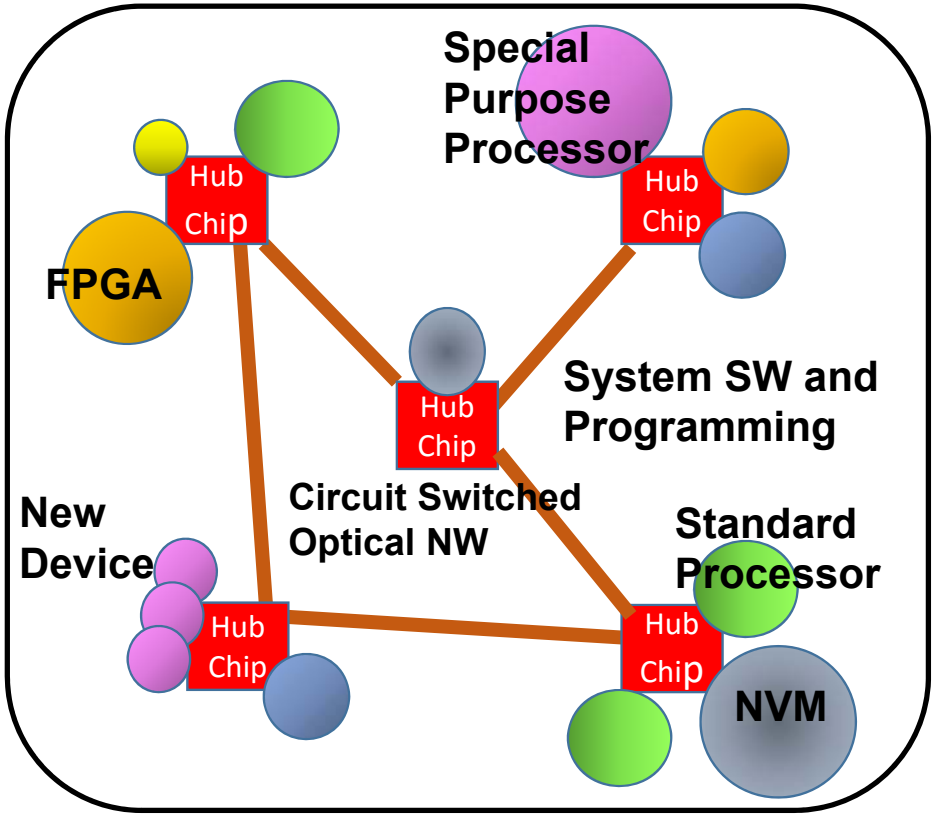


Important Properties of Post-Moore Architecture (Hideharu Amano, Keio Univ.)

- Able to connect a variety of new devices
- And integrate them into a system
- Proposal: “Super Building Block Architecture”
- HubChip: flexible HW reconfiguration including SW and Firmware
  - Inductive TCI（Through Chip Interface)
  - Accelerator-in-Switch (c.f. Taisuke’s talk)
  - Reconfigurable switches, FPGA, CGRA

Super Building

Holistic Control of Component Power

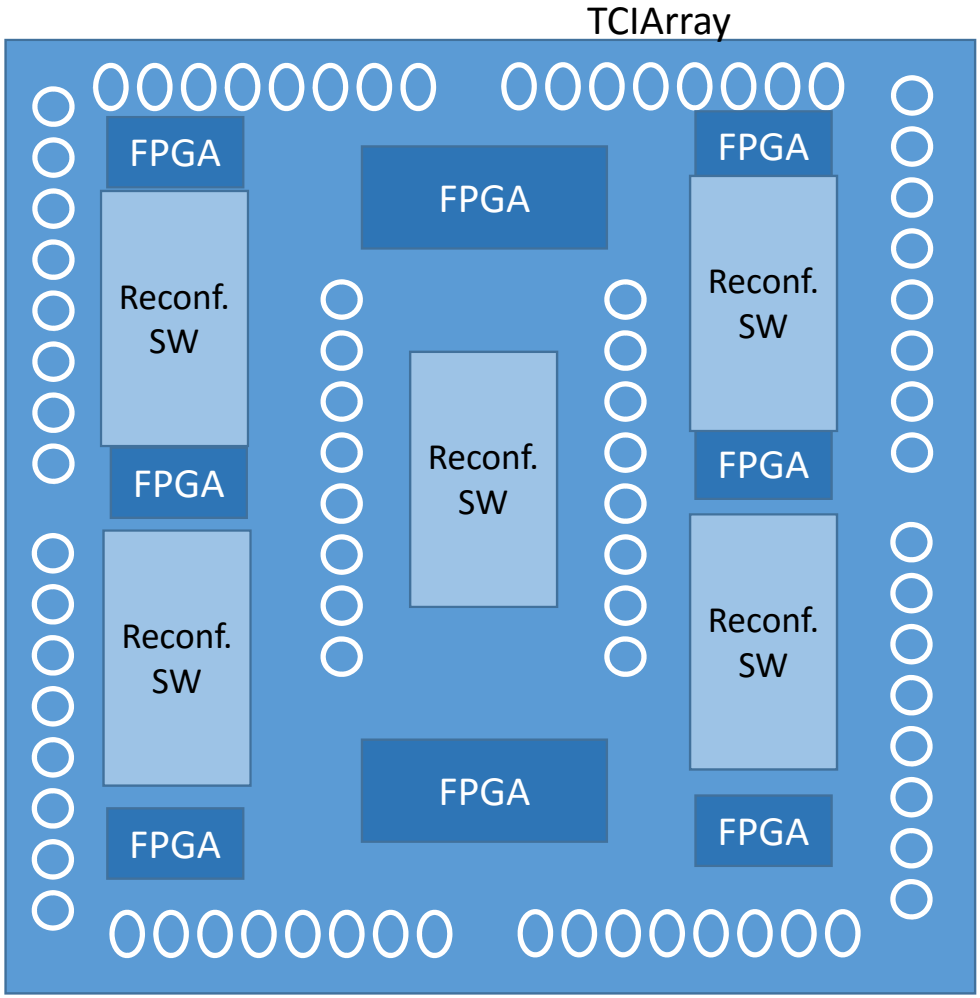


System View on "Post-Moore" Architecture

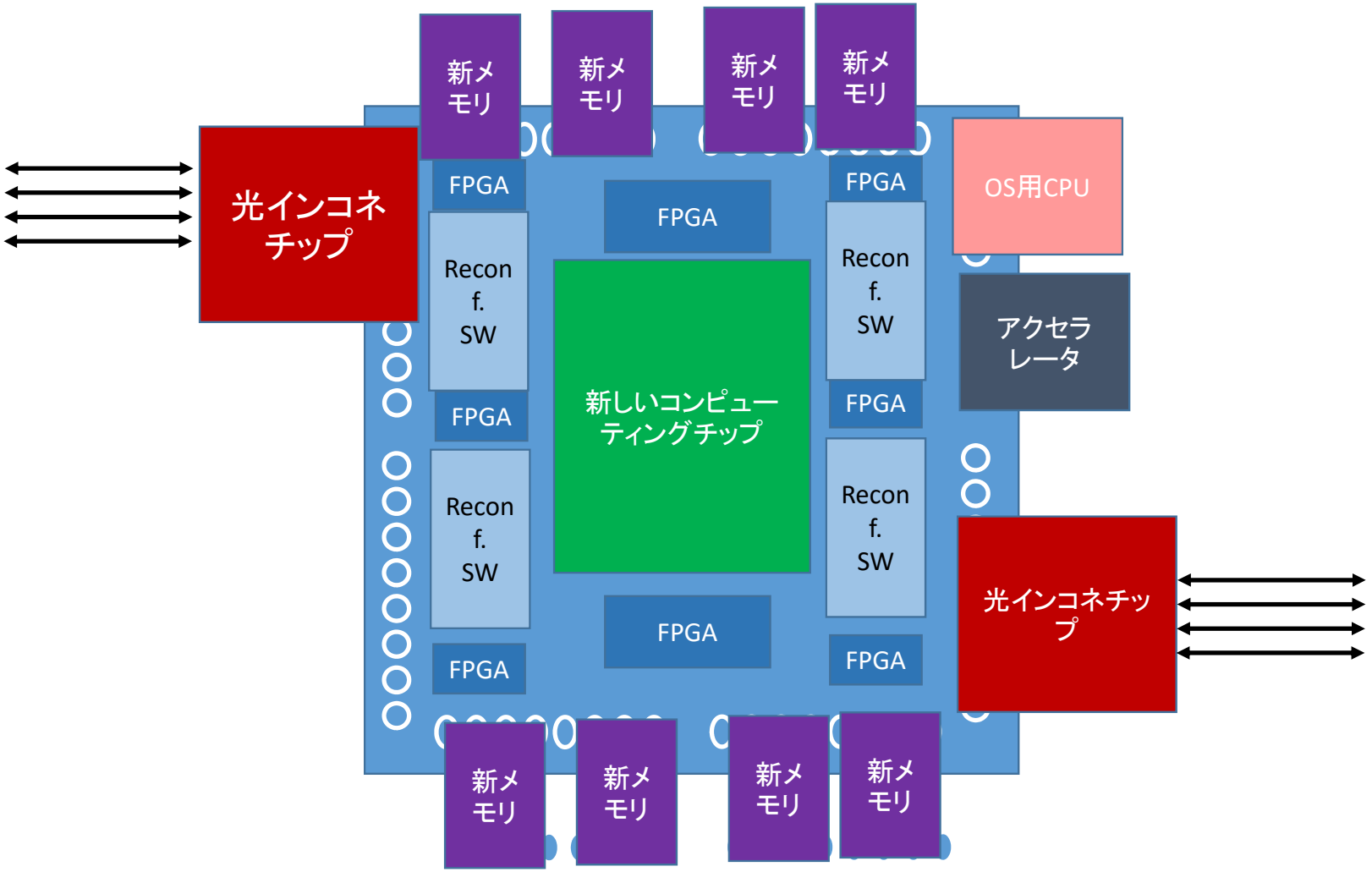
Not just a new device, but focus on how they are interconnected, and integrated as a system controlling their power

A Hub architecture that employs Inductive (3D) TCI and programmable FPGA+Switch

Hub Base Chip = Inductive TCI + Reconfigurable Swtich + FPGA

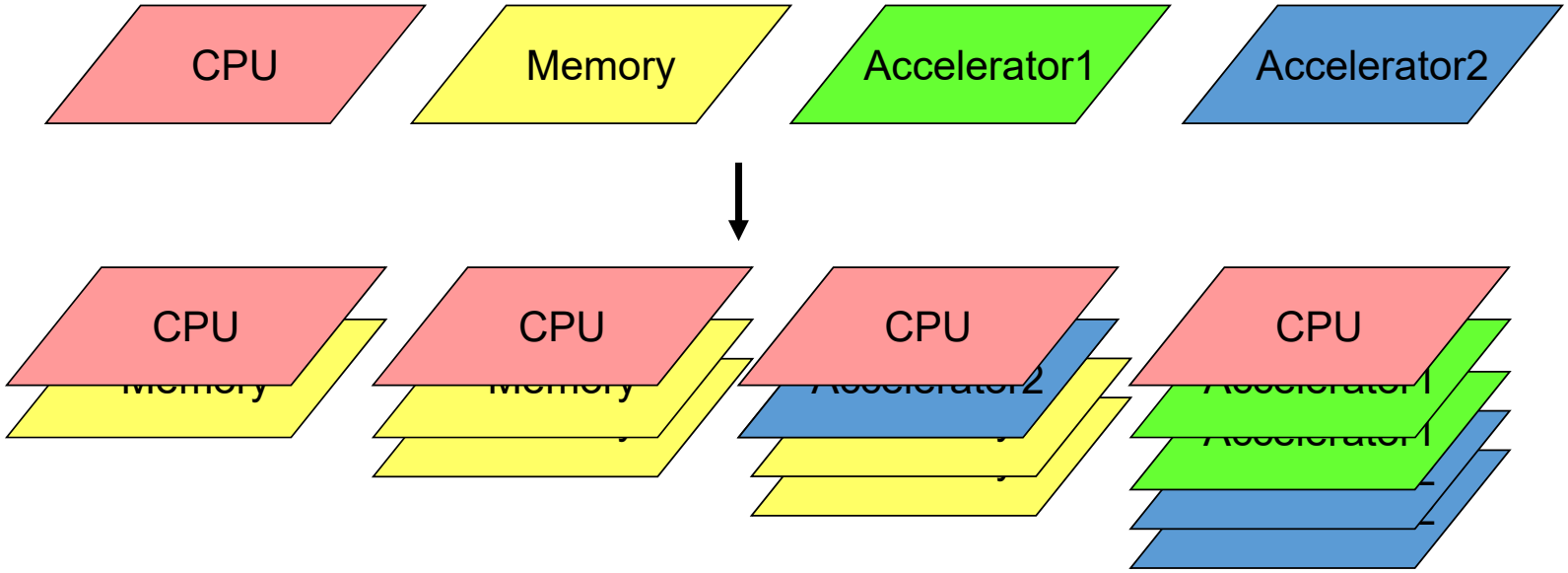


A new reconfig device based on TCI arrays and reconfigurable switches



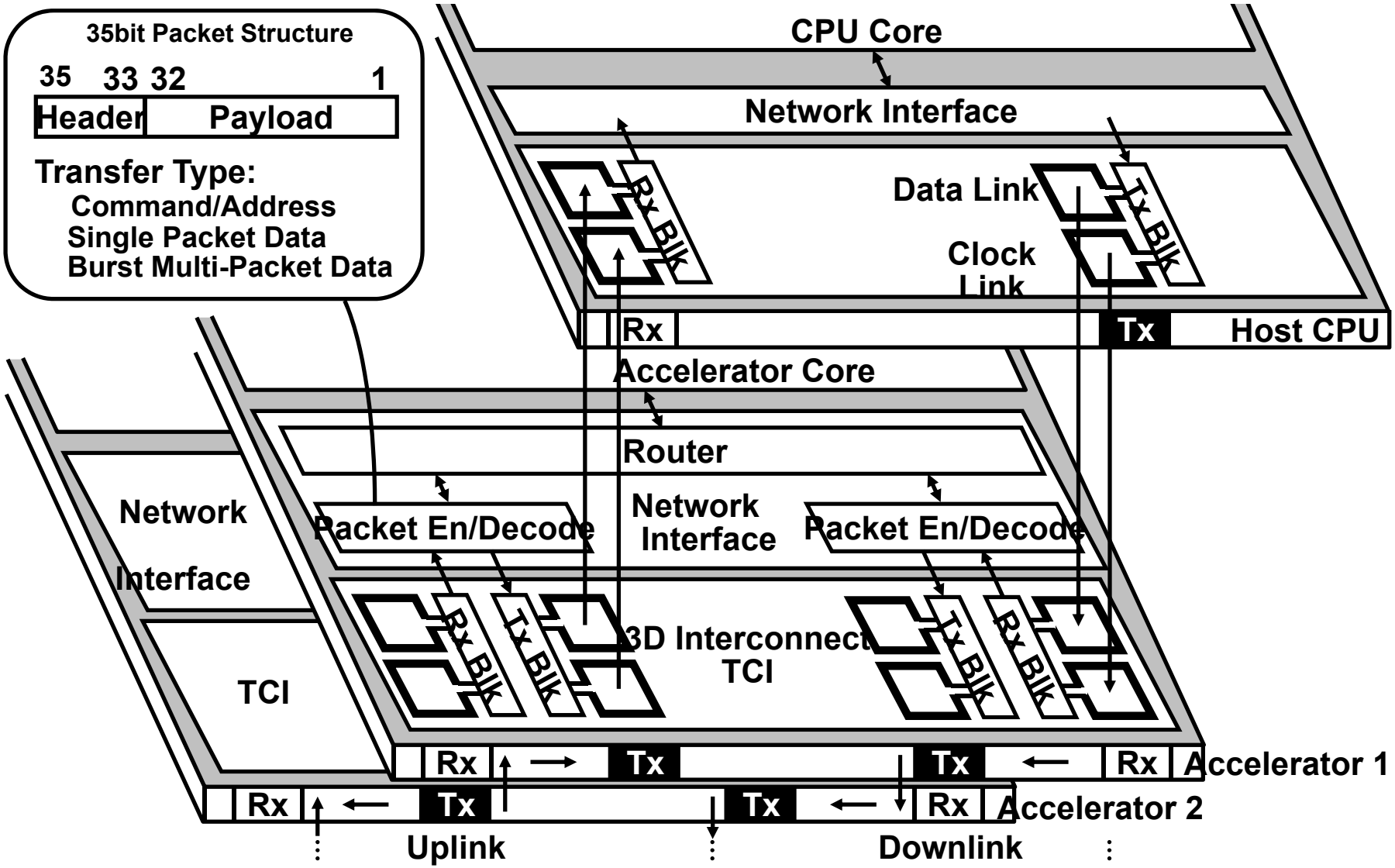
ドーターチップ接続のイメージ

Base Technology : Inductive TCI and Building Block Computing System



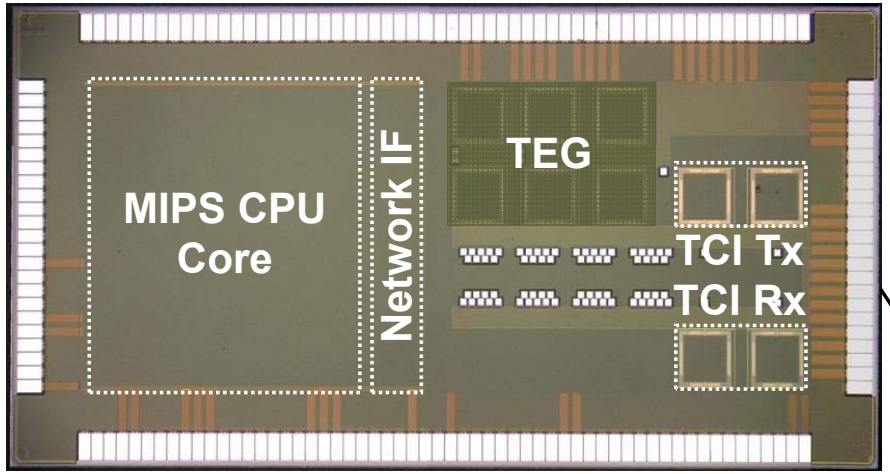
1 step: Various Combination can be done after chip-fabrication
2 step: Chips can be replaced by users → Field Stackable

Key Technique: Inductive Coupling Through-Chip Interface（TCI)

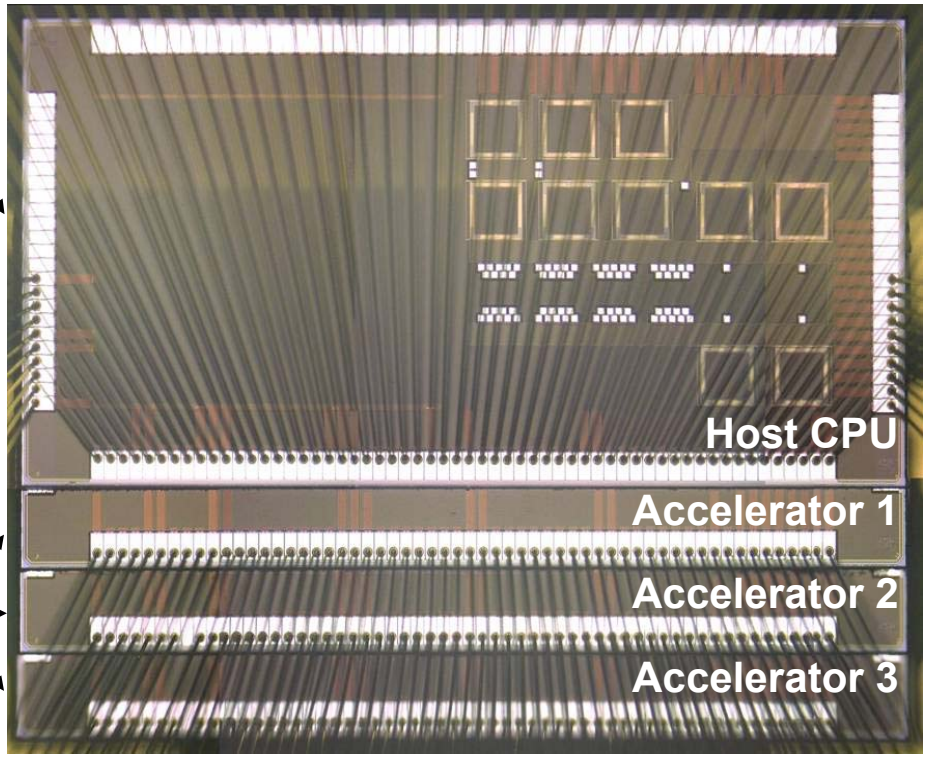


Block diagram of scalable 3D NoC using inductive-coupling ThruChip Interface (TCI).

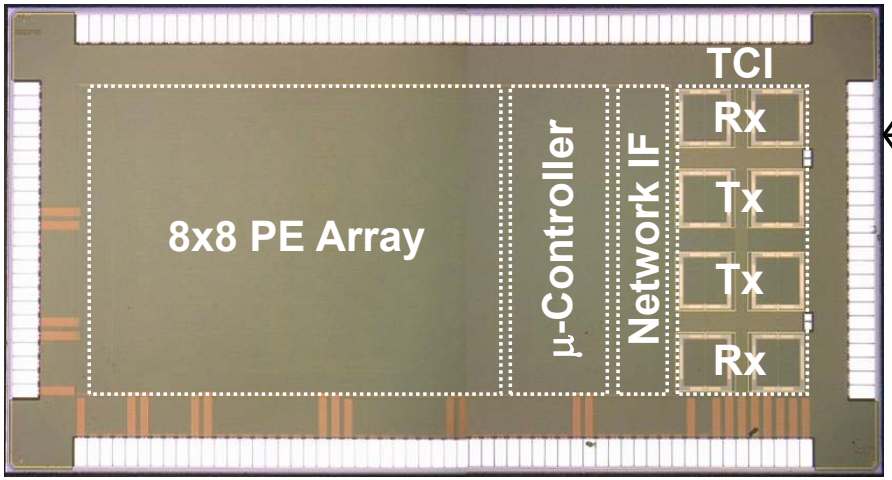
8Gbps 10mW/link 10^-9 error rate



Host CPU Chip



Host CPU + Accelerator x3 Chip Stack
Fabricated in 65nm CMOS

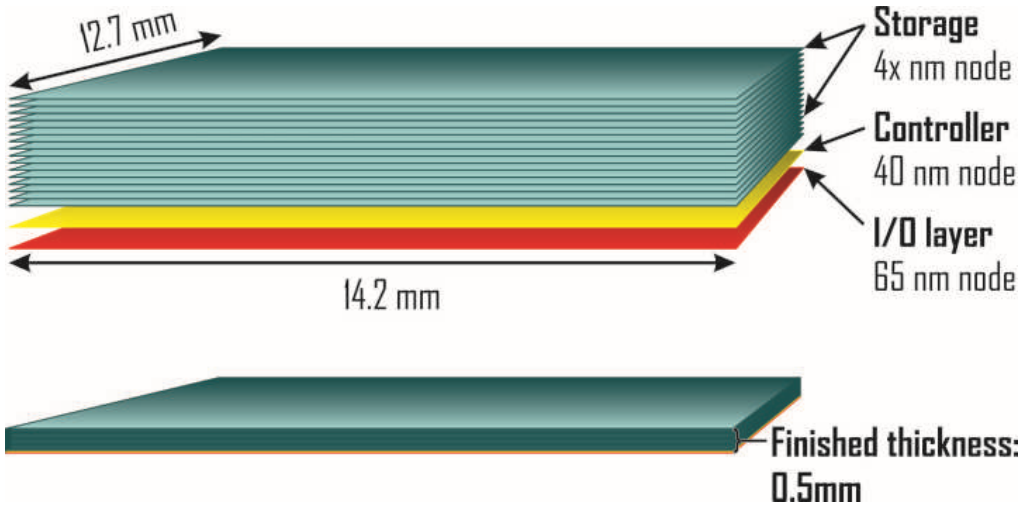


Accelerator Chip

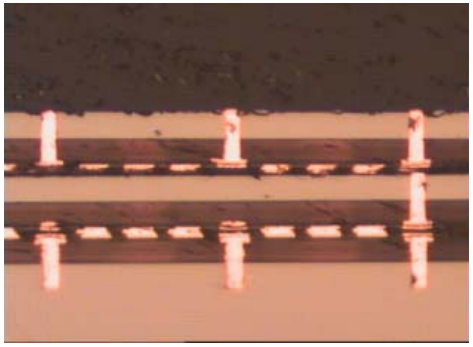
Microphotograph of stacked test chips.

Example Innovation: Tungsten TSV at 2um ultra fine pitch with die thinning by Tezzaron Semiconductor

- Suppose 4TF SFP @ 7nm, 16TB/s internal chip BW vs. 200GB/s external chip mem BW => 80 times speedup!
- High-density, high-signaling TSV challenge
  - Wide I/O 2 1024 bits 1 Ghz -> 2~3 Ghz
  - We need 128,000 bits @ 1Ghz !
  - 10 micron TSV estimation
    - 400 x 400 TSVs on 20mx20m chip -> 50 micron spacing
    - With tungsten TSVs the chip area is negligible

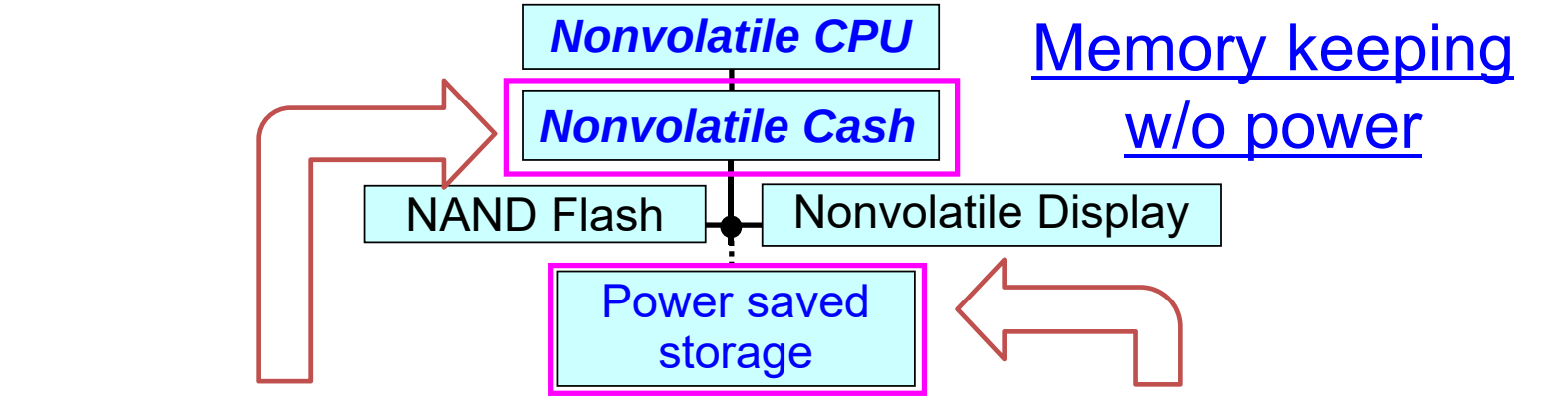


Many-layer stacking via aggressive wafer thinning and self-diagnostics



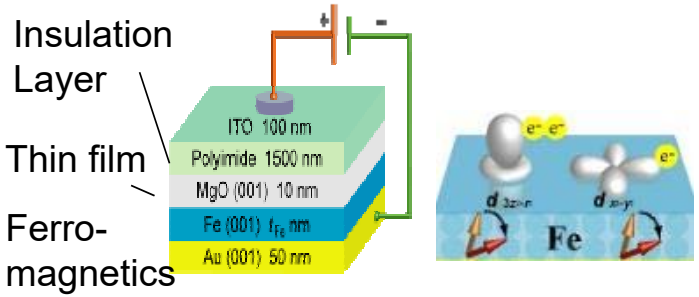
Source: Tezzaron website
http://www.tezzaron.com

Voltage-controlled Nonvolatile Magnetic RAM



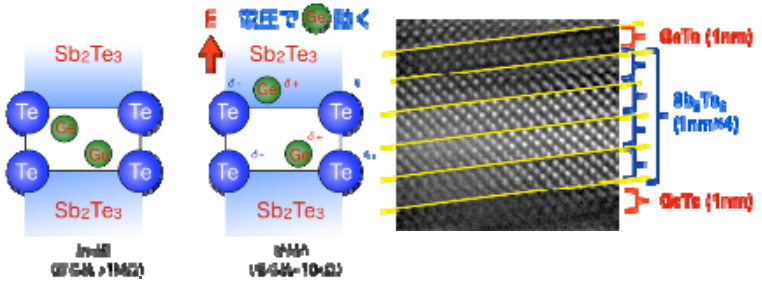
Memory keeping w/o power

Voltage Controlled Spin RAM



- voltage-induced magnetic anisotropy change
- Less than 1/100 rewriting power

Voltage Controlled Topological RAM

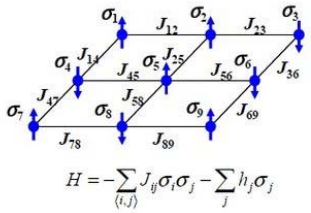


- Resistance change by the Ge displacement
- Loss by entropy: < 1/100

- Hitachi@ISSCC2015
  "An 1800-Times-Higher Power-Efficient 20k-spin Ising Chip for Combinational Optimization Problem with CMOS Annealing"
- Competitive to Quantum Annealing, room temperature, easy to scale
- Could be applicable to deep learning?

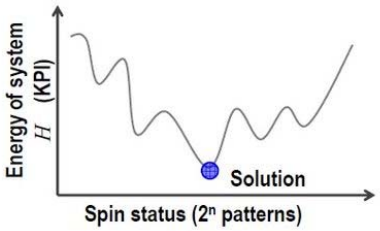
Computing with Ising model

- Ising model: expressing behavior of magnetic spins
- Using Ising model as natural phenomenon to map problems



$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j$$

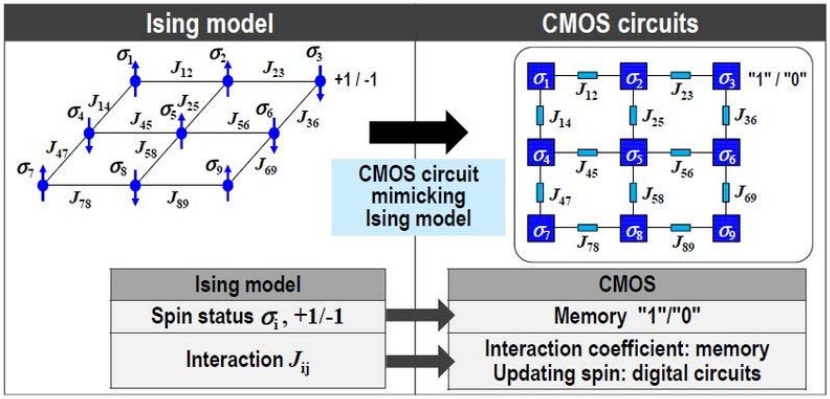
H: Energy of system
σ_i: Spin status (+1/-1)
J_{ij}: Interaction coefficient



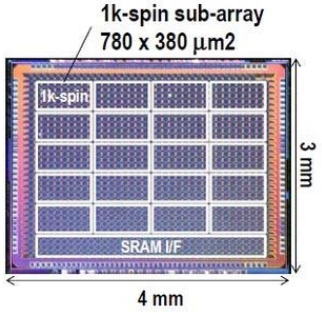
n: number of spins

CMOS Ising computing

- Mimicking Ising model with CMOS circuits
- Easy to manufacture, easy to use, good scalability



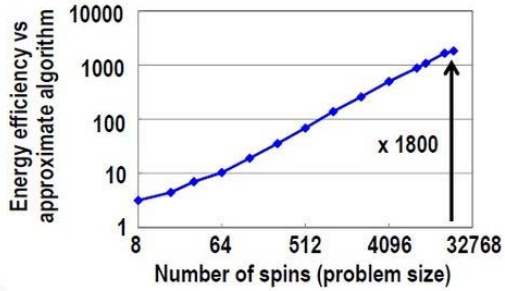
Fabrication results



Items	Value
Number of spins	20k (80 x 256)
Process	65 nm
Chip area	4x3=12 mm^2
Area of spin	11.27 x 23.94 =270 µm^2
Number of SRAM cells	260k bits Spin value: 1 bit Interaction factor: 2 bit x 5=10 bits External magnetic coefficient: 2 bits
Memory IF	100 MHz
Interaction speed	100 MHz
Operating current of core circuits (1.1 V) Do not include IO	Write: 2.0 mA Read: 6.0 mA Interaction: 44.6 mA

Measurement results of energy

- 1,800 times higher energy efficiency than conventional approximation algorithm on CPU



Conditions:
Randomly generated problems, energy for same preciseness solution
Ising chip: VDD=1.1 V, 100-MHz interaction, best solution among 10-times trial is selected.
Approximation algorithm: SG3(*) is operated on Core i5, 1.87 GHz, 10 W/core.

(*): Sera Kahruman et al., "On Greedy Construction Heuristics for the Max-Cut Problem," International Journal on Computational Science and Engineering, Volume 3, Number 3/2007, pp. 211-218, 2007.

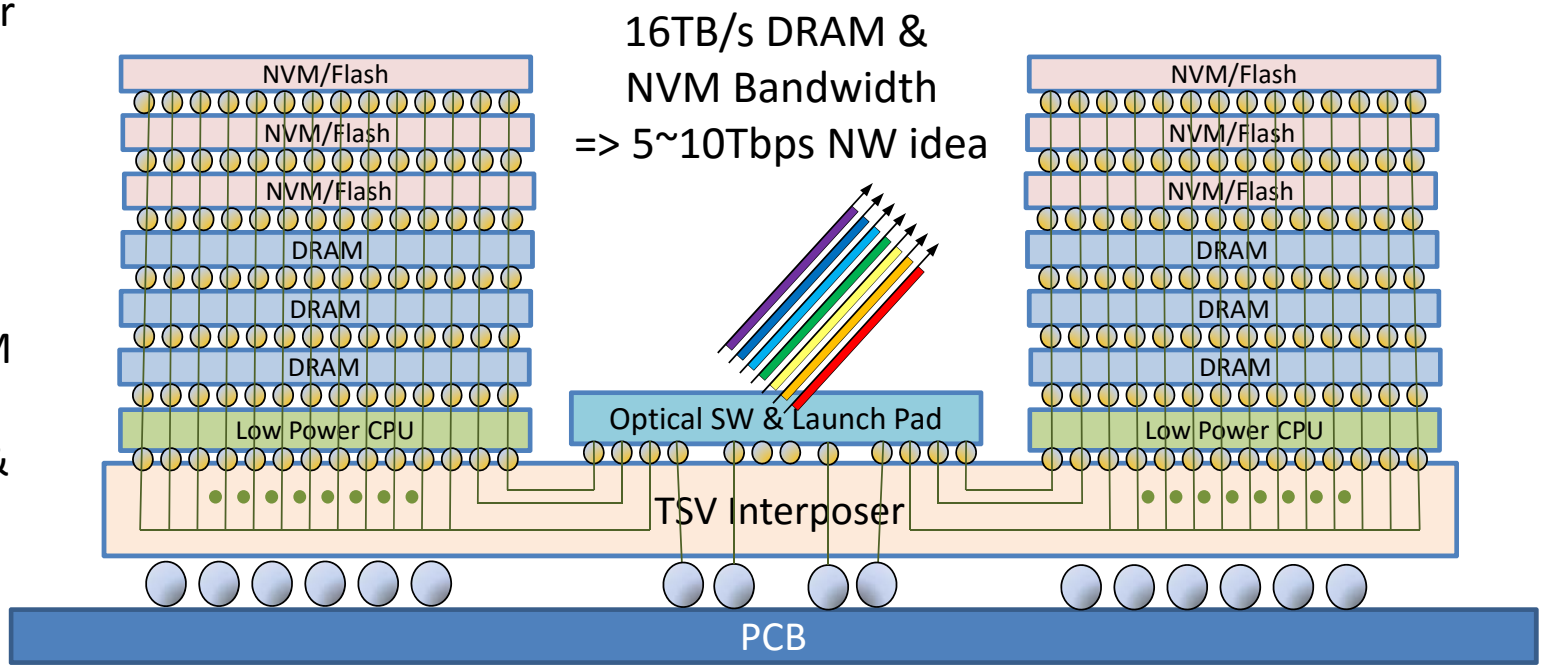
Strawman BYTES-Oriented Post-Moore Architecture

Low voltage & power CPU for direct stacking and large silicon area

Domain-specific hetero- and customizable processor configurations, including PIM

Extreme multi-layer DRAM & NVRAM stacking via high density tungsten TSV

Direct WDM optics onto Interposer



Direct Chip-Chip Interconnect with DWDM optics
Low Power Processor allows Direct 3D Stacking
Configurable Low-power CPU

Optical Network Technology for Future Supercomputers & IDC

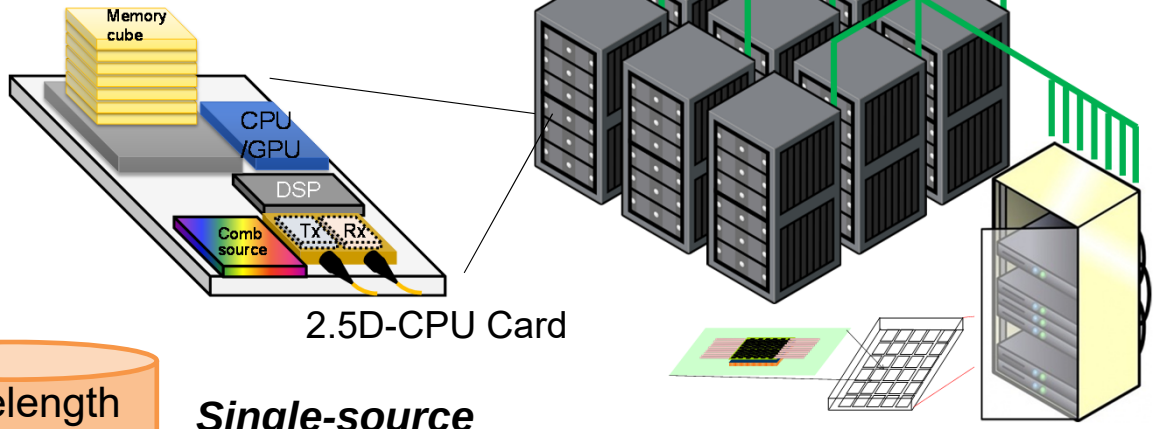
AIST IMPULSE

- Large-scale silicon photonics based cluster switches
- DWDM, multi-level modulation, highly integrated "elastic" optical interconnects
- Ultra-low energy consumption network by making use of optical switches

SC14 AIST Booth #2531

DWDM, multi-level modulation optical interconnects

Datacenter server racks



2.5D-CPU Card

Silicon photonics cluster switches

- Ultra-compact switches based on silicon photonics
- 3D integration by amorphous silicon
- A new server architecture

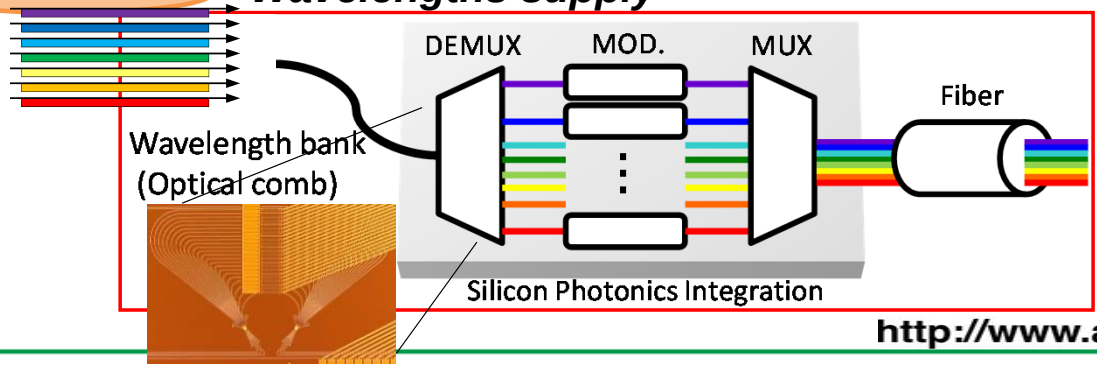
Current electrical switches:
～1pbps
↓
～500Pbps

No of λs	Order of mod.	Bit rate
1	1	20 Gbps
4	8	640 Gbps
32	8	5.12 Tbps

Current state-of-the-art Tx
100Gbps
↓
～ 5.12Tbps

Wavelength Bank

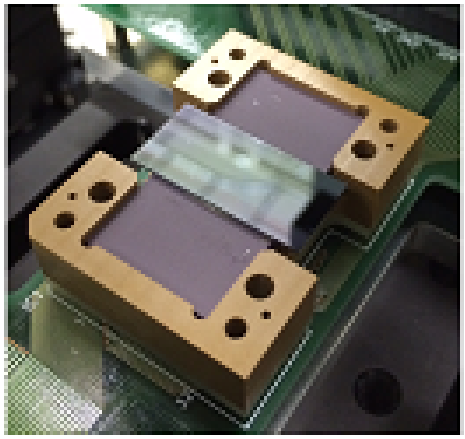
Single-source Wavelengths supply



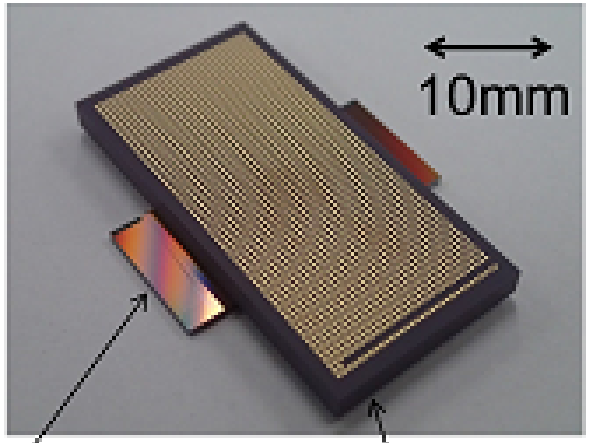
32 x 32 Optical Circuit Switch (Courtesy AIST)

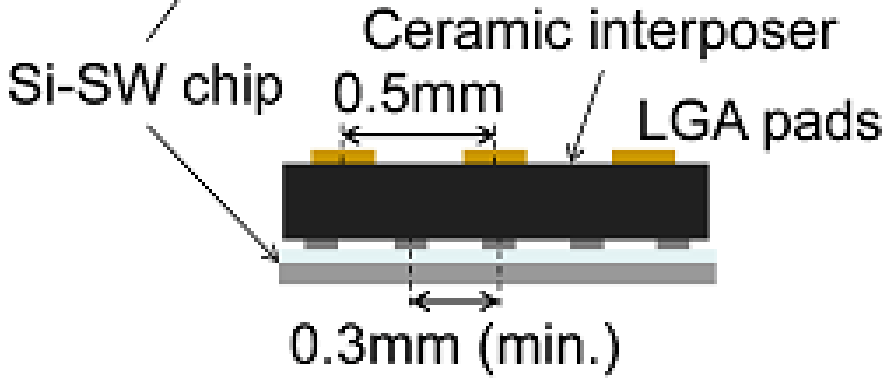
- Ceramic LGA interposer with 0.5-mm pitch
- Flip-chip bonding with Au bumps and non-conductive paste
- LGA socket to contact PCB

LGA socket



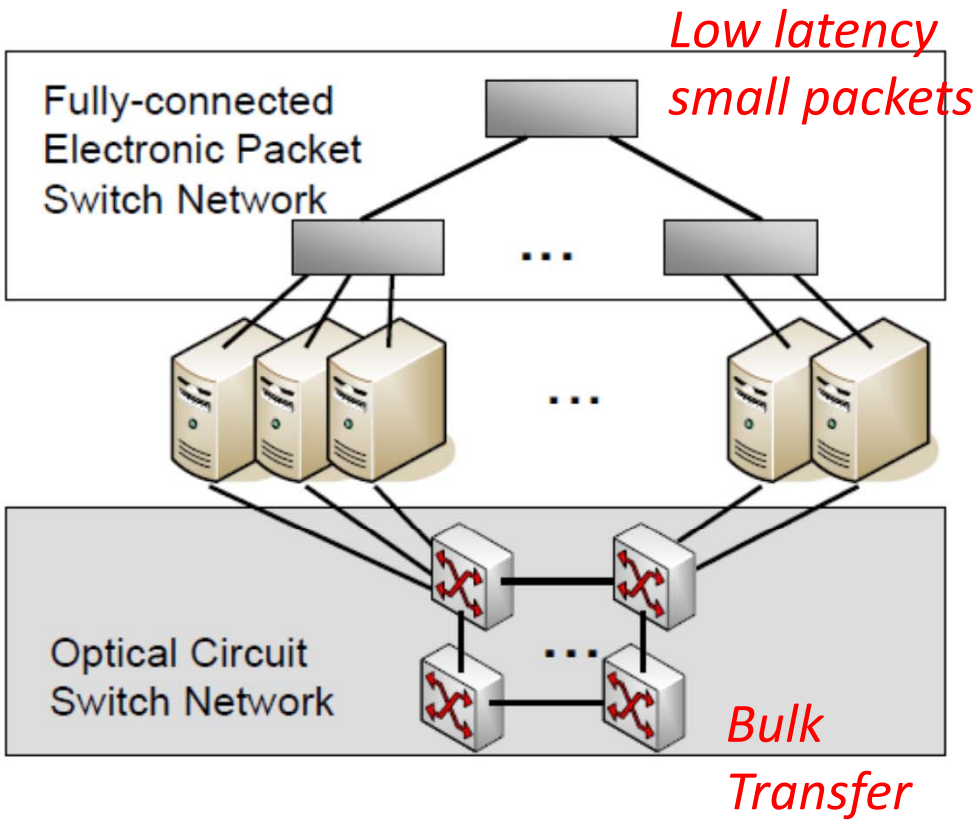
After FC bonding

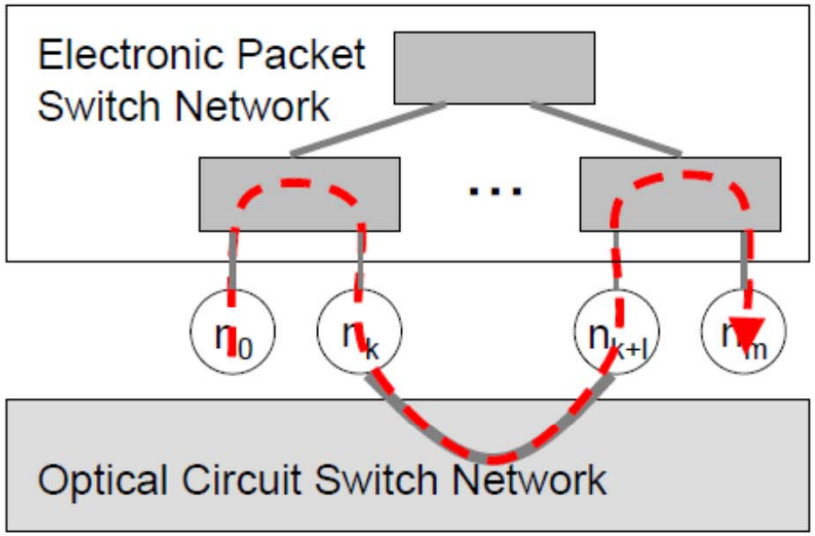




Hybrid Electro-Optical Network w/shortcuts [Takizawa&Matsuoka LSPP07]

"Locality Aware MPI Communication on a Commodity Opto-Electronic Hybrid Network"





EO-OE Shortcut Forwarding of Messages Across Switches

"We should *optimally* design future SCs like Cell Phones or even human brains so that we *never repair* them, but simply work around"

Image: Illustration by Mirko Ilic

SC14 Wed 2:00PM

Fail-in-Place Network Design

[SC14, Domke, Hoefler, Matsuoka]

What if we never repaired network failures? Will intelligent routing compensate?

All investigated routing algorithms show limitations

- Fat-tree, UpDown, DOR, Torus2QoS
- MinHop, SSSP, DFSSSP, LASH

Topology-aware routings

- High throughput decrease possible with small failure percentage
- Fail to route highly damaged netw.
- Routes not always DL-free (DOR)

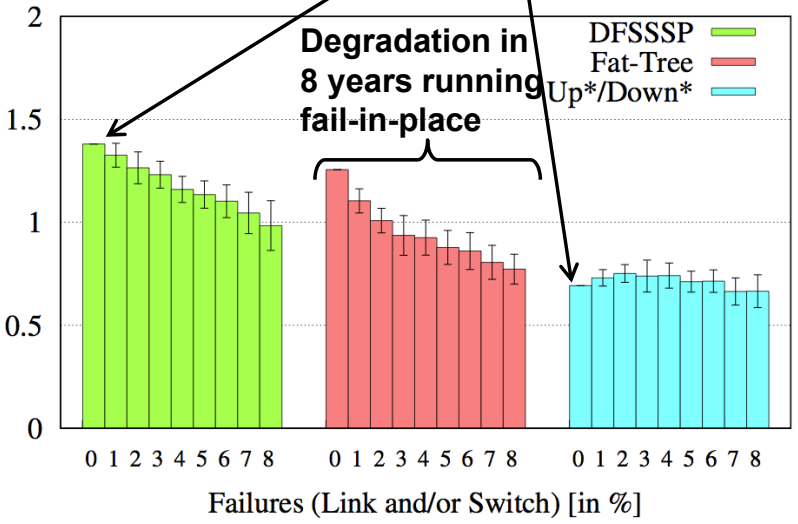
Topology-agnostic routing algorithms

- Ignore deadlocks (MinHop, SSSP)
- Deadlock-avoidance via VLs can be impossible for large scale netw.

TABLE II. INTERCEPT, SLOPE, AND R^2 FOR TSUBAME2.0

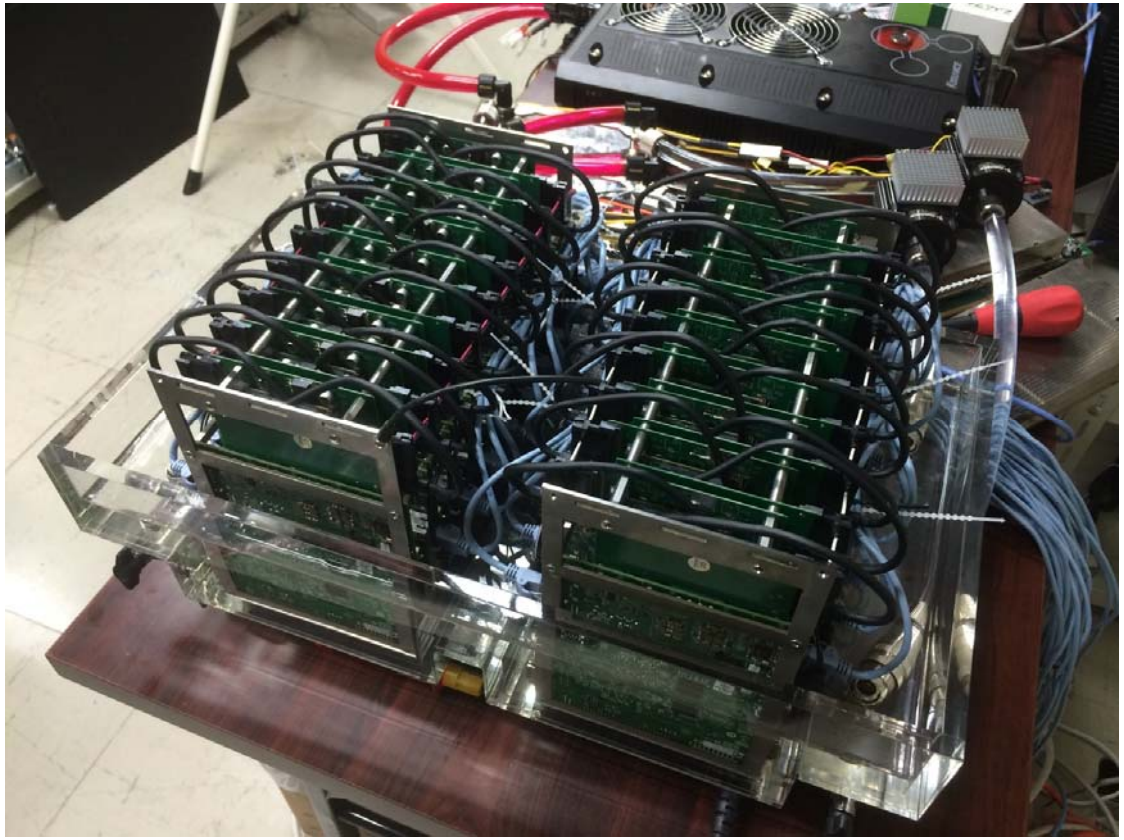
HPC system	Routing	Intercept [in Gbyte/s]	Slope	R^2
TSUBAME2.0	**DFSSSP**	1,393.40	-1.33	0.62
	Fat-Tree	1,187.19	-1.48	0.66
	Up/Down**	717.76	-0.08	0.01

Changing from Up*/Down* (default) to DFSSSP routing on TSUBAME2.5 improves the throughput by 2.1x for the fault- free network and increases TSUBAME's fail-in-place characteristics.



Fail-in-Place Network Design is possible! (but we have to improve the routing)

GoldenBox "Proto1" (NVIDIA K1-based) Demo at SC14 Tokyo Tech. Booth #1857

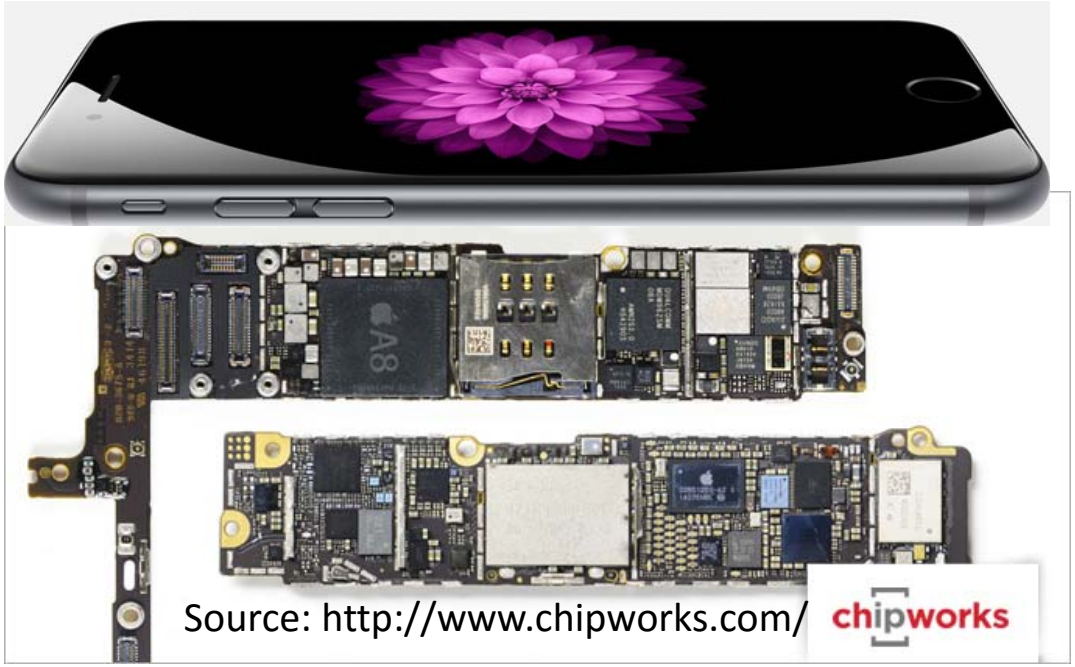


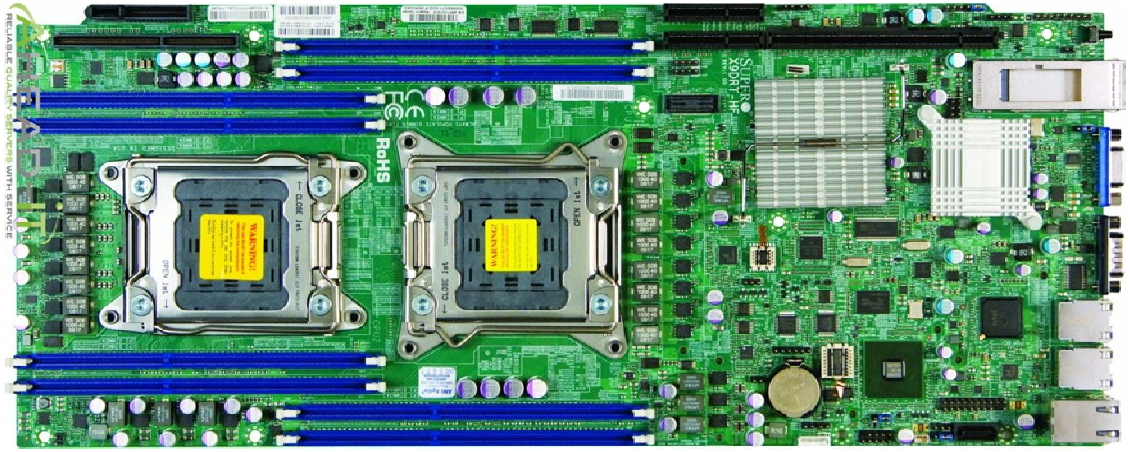
- 36 Node Tegra K1, ~11TFlops SFP
- ~700GB/s BW
- 100-700Watts
- Integrated mSata SSD, ~7GB/s I/O
- Ultra dense, Oil immersive cooling
- Same SW stack as TSUBAME2

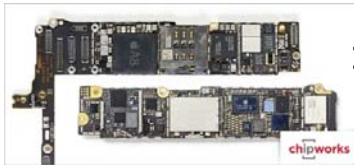
2022: x10 Flops, x10 Mem Bandwidth, silicon photonics, x10 NVM, x10 node density, with new device and packaging technologies

2013 TSUBAME-KFC 4GPUnode

VS.







2025 Post-Moore Node
5nm SoC technology
4 Teraflop CPU+GPU SFP
16TB/s fast memory
>1 Terabyte NVM
1TB/s NW (to neighbor node and to long-range optics)
15~25Watts
$200 – exploits ecosystem
Runs today's software!

Innovation 2025 K-in-a-Box (Golden Box)
Post-Moore Architecture

1/1000 Size, 1/150 Power, 1/500 Cost, x20 DRAM+ NVM Memory

20 Petaflops, 20 Petabyte Hiearchical Memory (K: 1.5PB), 10K nodes

1TB/s Interconnect (>100Pbps Bisection BW)

(Conceptually similar to HP “The Machine”)

Datacenter in a Box

Large Datacenter will become “Jurassic”

Towards Understanding the Performance of FPGAs using OpenCL Benchmarks [HiPEAC Reconfigurable Computing WS 2015]

Hamid Zohouri (Tokyo Tech), Naoya Maruyama (Riken AICS), Satoshi Matsuoka (Tokyo Tech), Motohiko Matsuda (RIKEN AICS)

In collaboration with:
Aaron Smith (Microsoft Research),

Supported by Altera

Customized Computations w/FPGA Exploiting Dark Silicone

- FPGAs
  - Hardened units + reconfigurable routing
  - Middle-ground solution to specialization and generality
- Become part of the HW/SW ecosystem of data centers?
- The questions
  - Is it really better than GPUs/Xeons/Xeon Phis?
    - Dense regular computations → Why not GPUs?
    - Irregular sparse computations → Reconfigurability does not help?
    - Lots of successful case studies in HDL → OpenCL?
  - How can it be programmed?
    - How effective is OpenCL for FPGAs?
    - Higher level programming models?

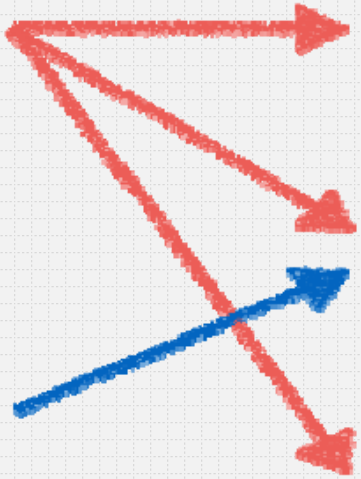
Evaluating FPGAs with OpenCL

- Rodinia benchmark suite [Che et al. 2009]
  - Consists of 23 benchmarks, each with OpenMP, CUDA and OpenCL versions
  - Covers the Berkeley Dwarfs
- Evaluation approach
  - Use the original OpenCL version as basisline
  - Some modifications in makefile and host code for compatibility with FPGAs
  - We assume the benchmarks are already well-optimized for CPU and GPU
  - NW, Hotspot, and Pathfinder have been evaluated and optimized

Parallelism in Altera OpenCL

Explicit:
Thread/SIMD parallelism

Implicit:
Pipeline parallelism



- **Inter pipelines**
  - Configurable number of duplicated pipelines
- **Intra pipeline**
- **SIMD**
  - Instantiate SIMD units base on user direction (attribute **`num_simd_work_items`**)

Performance Evaluation

Hardware

- FPGA: Terasic DE5-Net Board
  - Stratix V 5SGXA7
  - ALM: 234,720
  - Register: 938,880
  - M20K: 2,560
  - DSP: 256
  - Altera Quartus v15.0.2
- GPU: Tesla K20C
  - 2496 CUDA cores @ 706 MHz
  - 5 GB GDDR5 memory
  - CUDA v7.5
- CPU: Intel Xeon E5-2670
  - 8-core Sandy Bridge-EP @ 2.6 GHz
  - 32 GB DDR3 memory
  - GCC v4.9.2

Benchmark

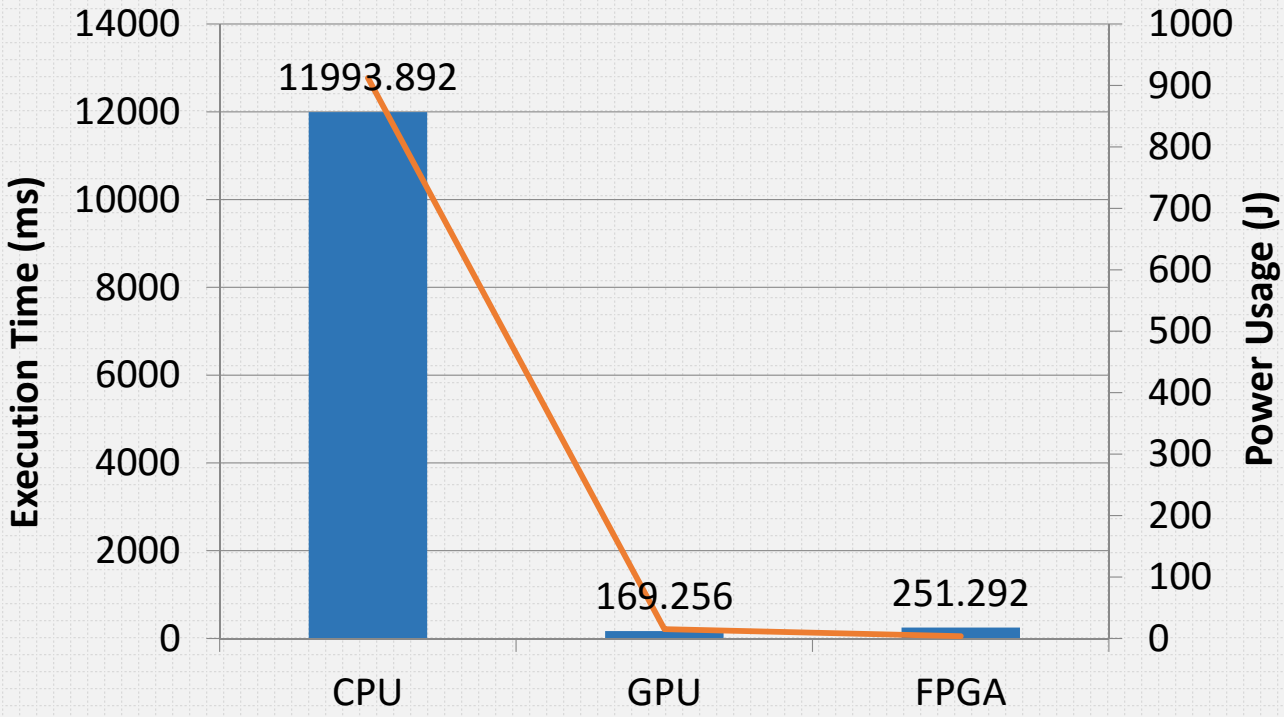
- Needleman-Wunsch (NW)
  - Integer dynamic programming benchmark
  - Data dependency on left, top left and top neighbors
- Hotspot
  - Single-precision 2D 5-point stencil computation
- Pathfinder
  - Integer dynamic programming benchmark
  - Data dependency on bottom, bottom left, and bottom right neighbors

Optimization Effects

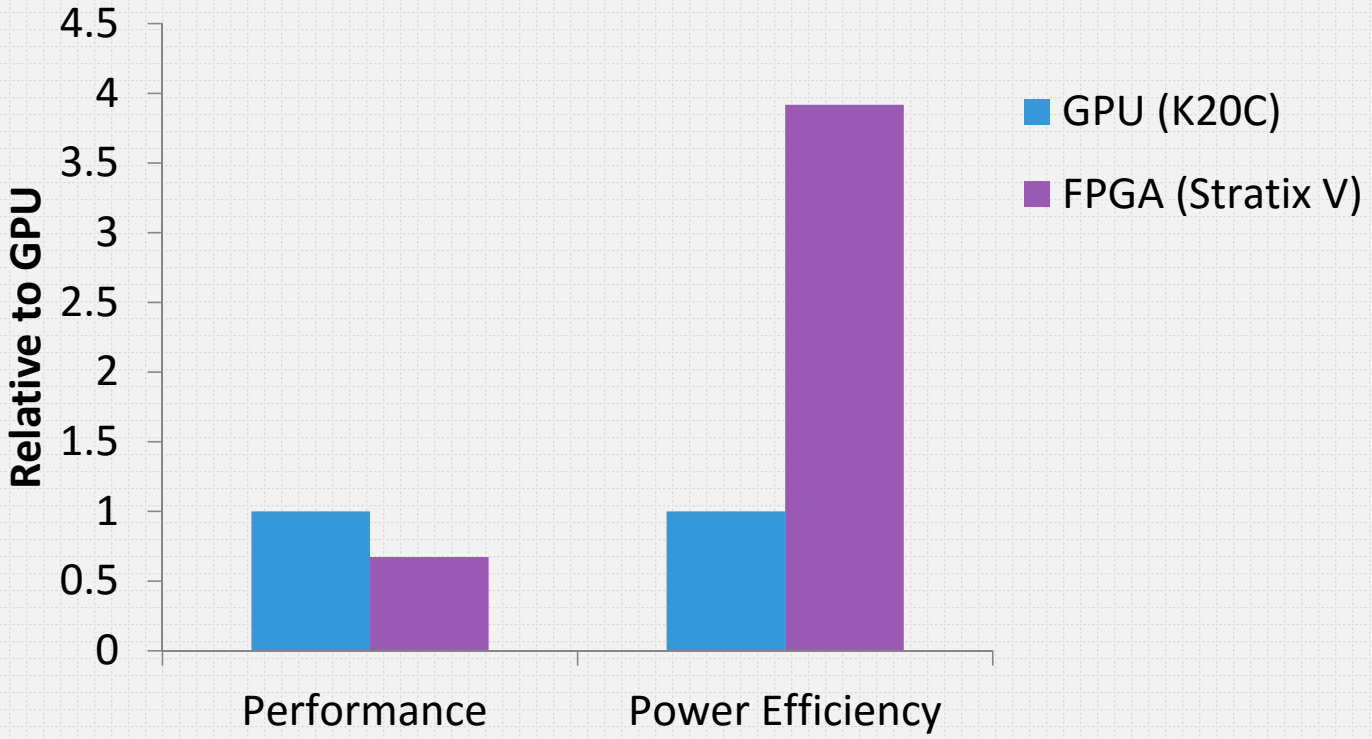
Type	Optimization	F_{max} (MHz)	Run Time (ms)	Power Dissipation (Watt)	Power Usage (J)
MT	None	277.2	16574	12.01	199.1
Pipeline	None	243.4	117523	10.59	1245.2
MT	Basic	194.7	2445	16.94	41.4
Pipeline	Basic	249.1	116457	9.93	1156.7
Pipeline	Advanced	148.0	251	15.44	3.8

Sliding window is 66x faster than baseline

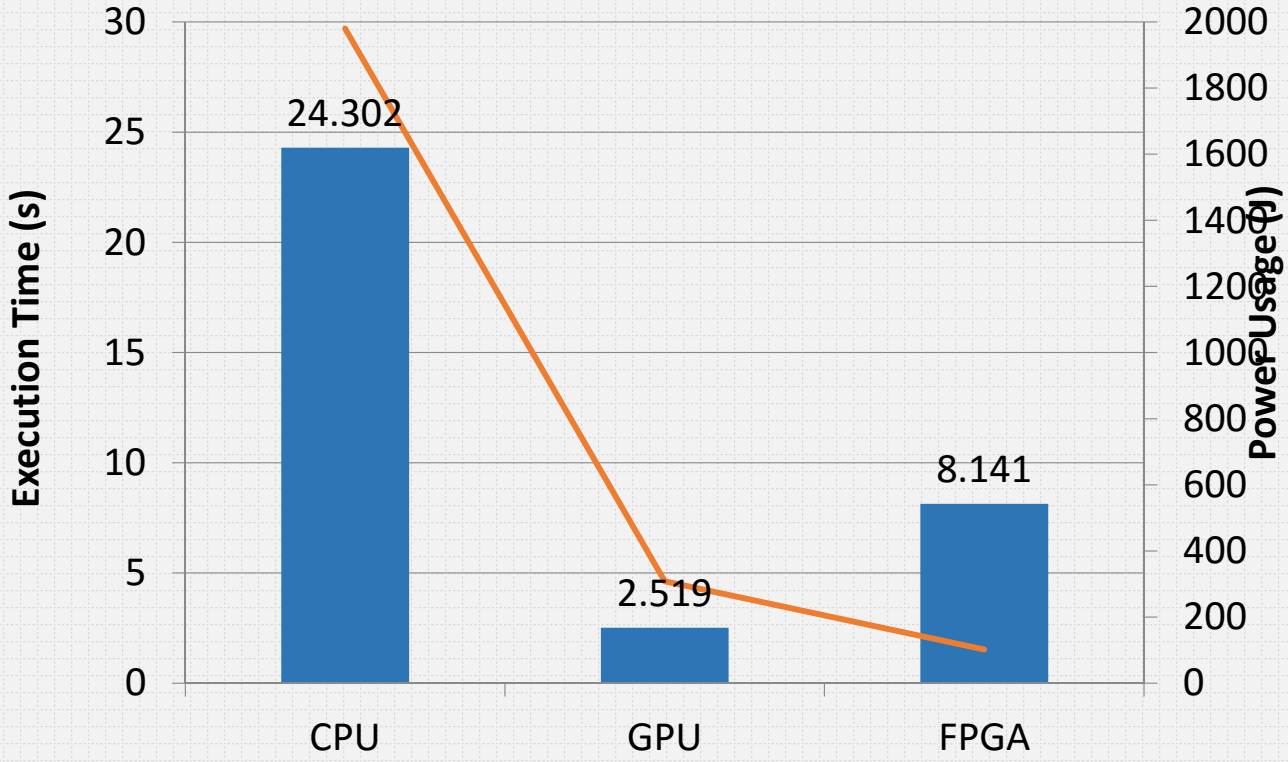
Comparison with CPU & GPU



FPGA vs. GPU


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Relative to GPU
Performance
Power Efficiency
GPU (K20C)
FPGA (Stratix V)

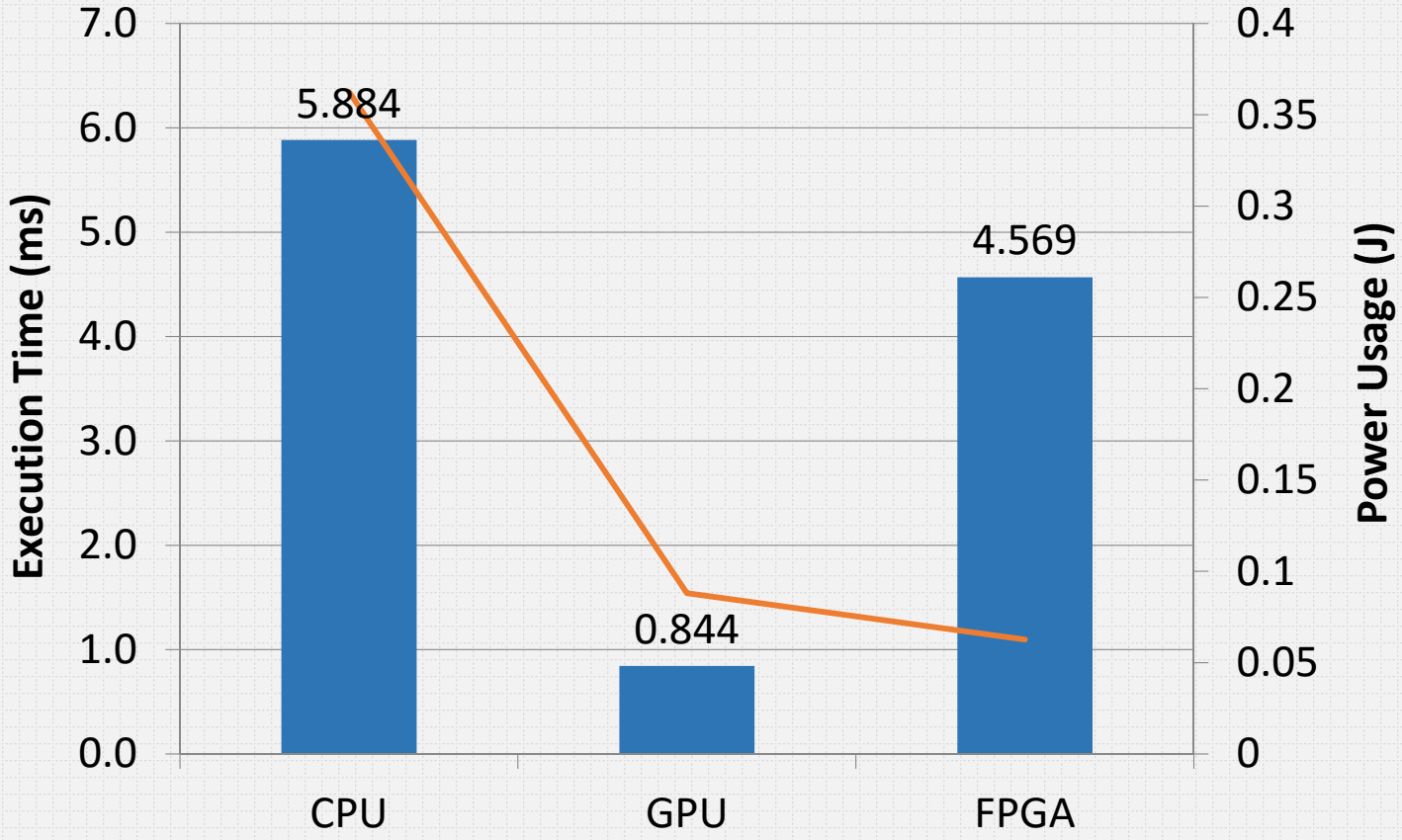
Hotspot



Time: 3x vs. CPU and 0.31x vs. GPU

Power efficiency: 19.4x vs. CPU and 3x vs. GPU

Pathfinder

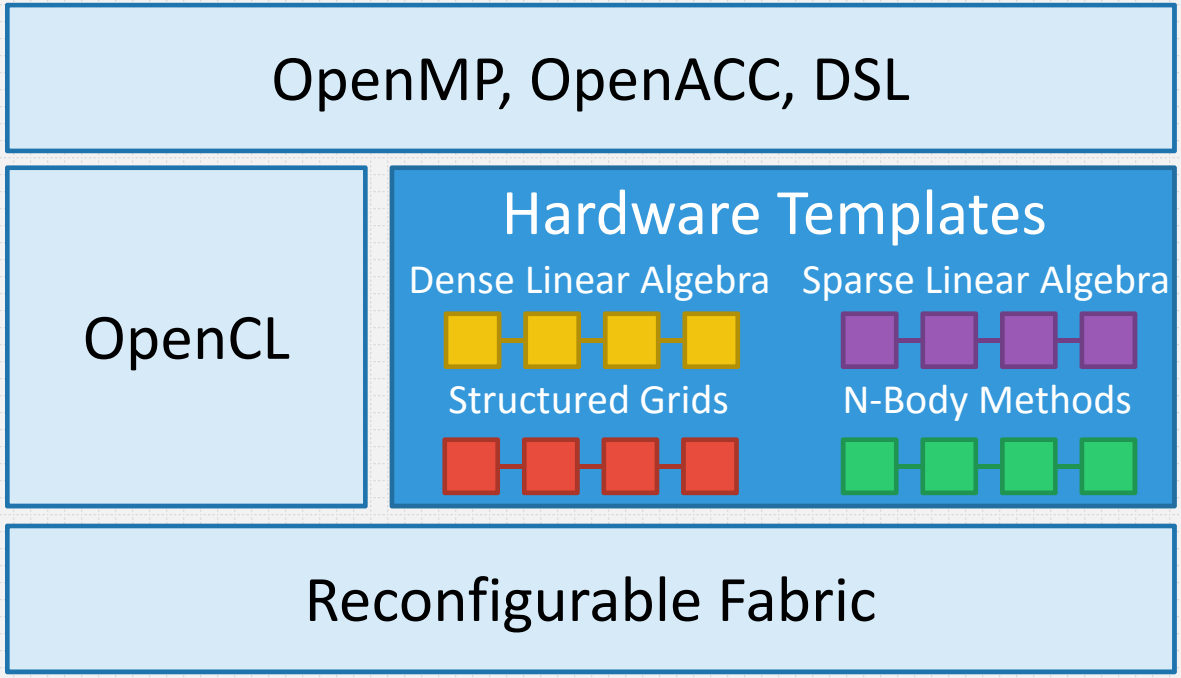


Time: 1.3x vs. CPU and 0.18x vs. GPU

Power efficiency: 5.8x vs. CPU and 1.4x vs. GPU

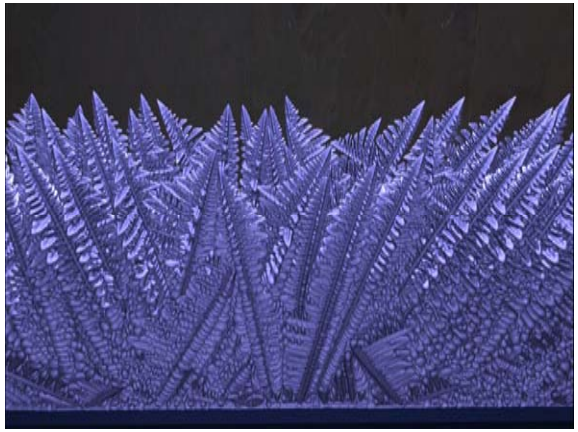
Towards Performance Portability

- Catalogue of reusable efficient hardware templates for common computation patterns
- Building blocks for implementing high Performance, high-Level programming Languages

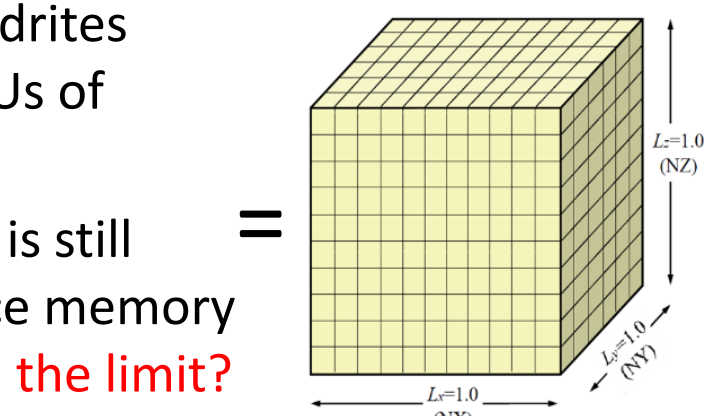


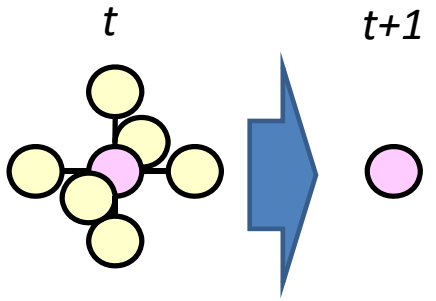
Communication Reducing Runtime for Stencil Simulations [Endo, Takasaki, Matsuoka ICPADS2015]

- CFD Simulations on GPU supercomputers
  - Stencil computations successfully accelerated by high FLOPS and high BW of GPUs
  - However, problem sizes (Bytes) are limited by device memory capacity

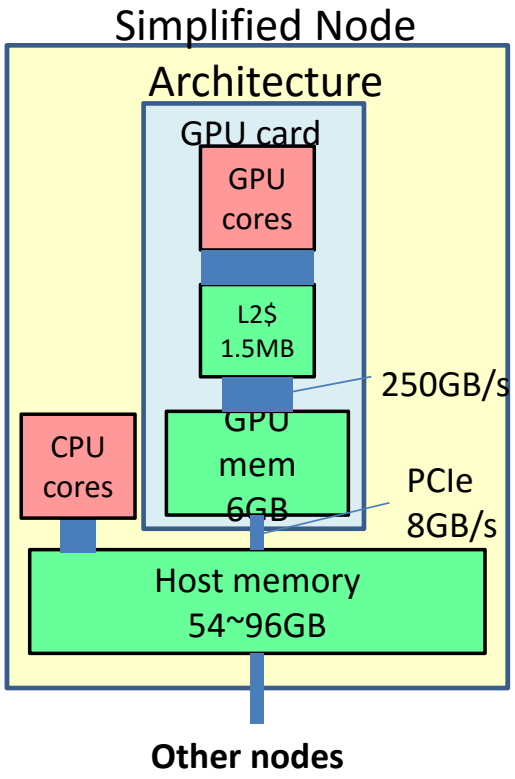


- Denderite Simulation by Shimokawabe, Aoki et al.
  - Multi-petascale Metallic Dendrites simulation, using > 4,032 GPUs of TSUBAME2/2.5
  - However, the simulation size is still limited by (aggregated) device memory capacity. How can we exceed the limit?



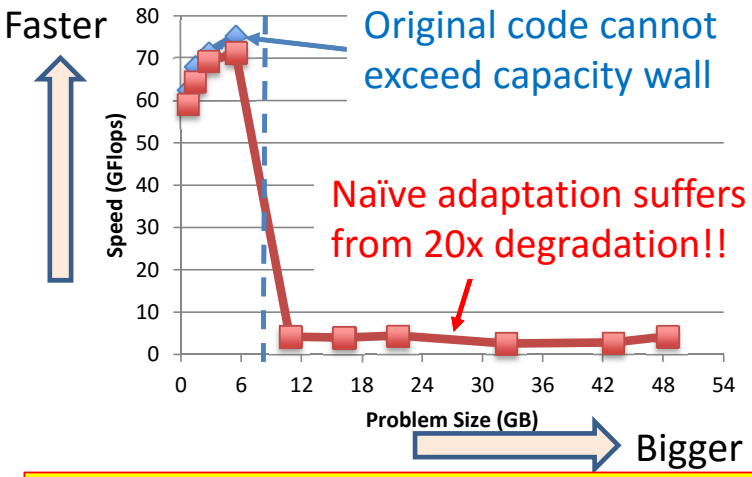


Two Issues for Extreme Big Simulations



Performance issue

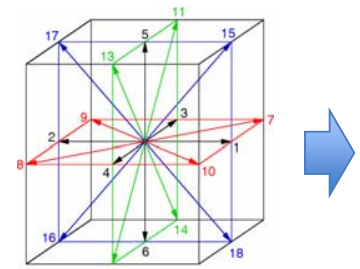
- Naïve usage of memory hierarchy causes 20x performance degradation
- Due to large PCIe costs

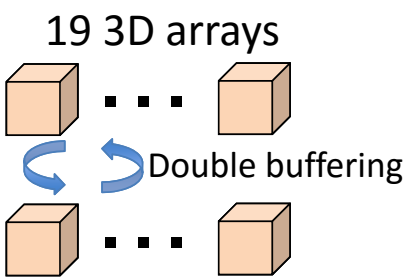


We need locality improvement / comm. reducing → *Temporal blocking* technique is adopted

Productivity issue

- Typically , temporal blocking introduces heavy code rewriting
- Original Wind simulation code is already complex
  - ~15,000 Lines of code with MPI/CUDA
  - LBM with hybrid boundary conditions (periodic and Neumann)

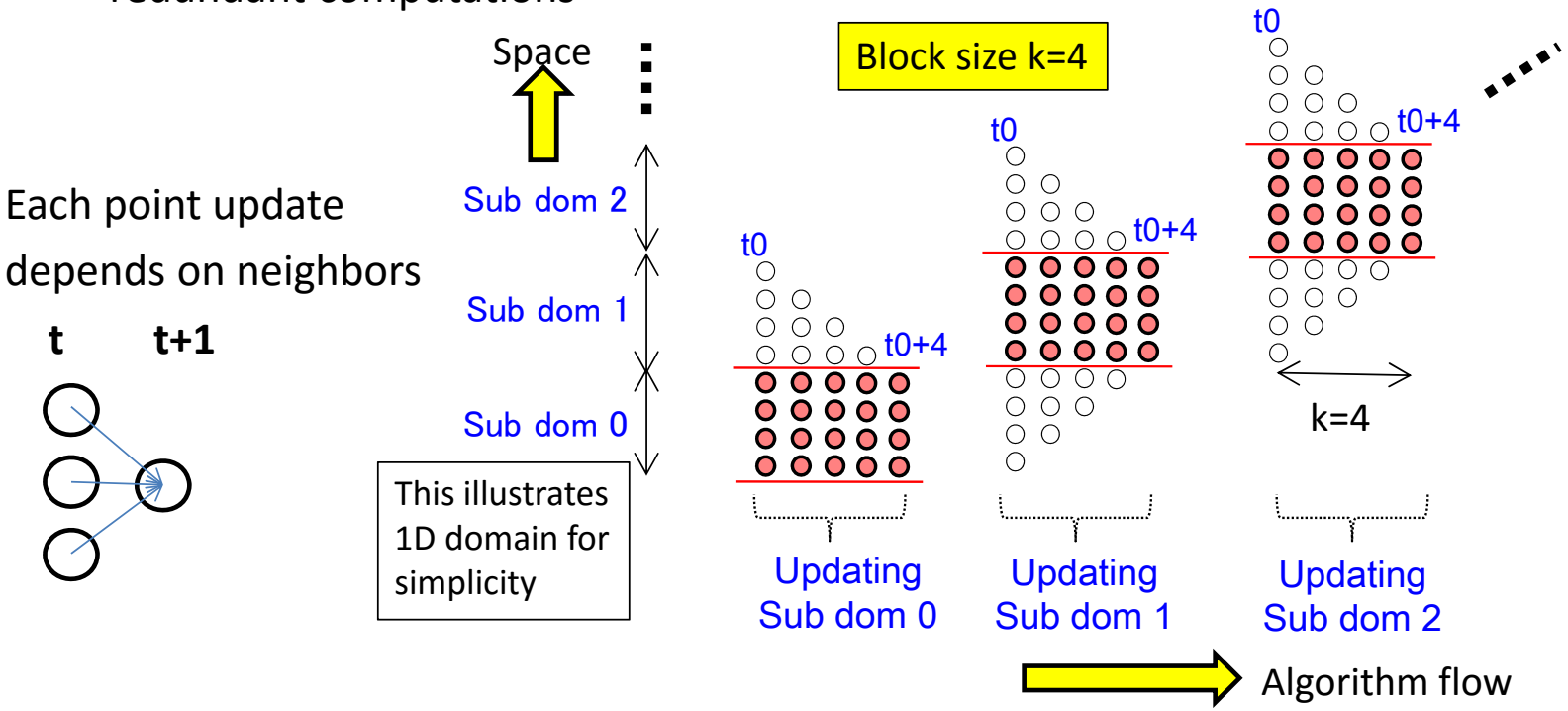

D3Q19 model



We reduce programming costs by using runtime library for memory swapping, HHRT

Temporal Blocking (TB)[Wolf 91]: Locality Improvement of Stencil Computations

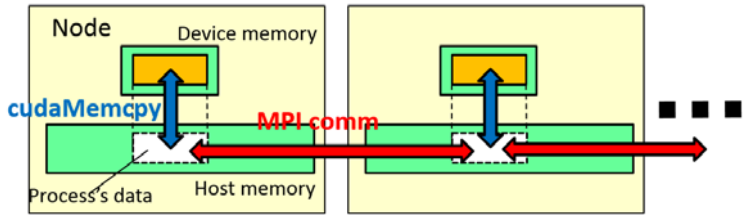
- The domain is divided into sub-domains
- When we pick up a sub-domain, we proceed its computation for several (=k) time steps at once
- This optimization introduces
  - re-structuring of loop structure
  - redundant computations



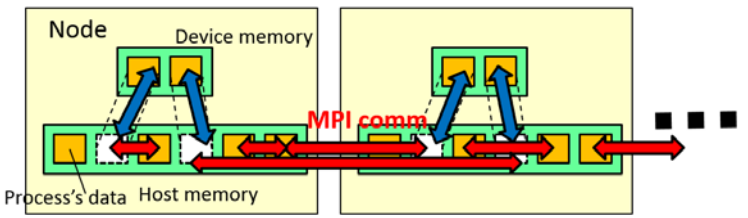
Using HHRT Library for Reducing Code Re-structuring Costs

- HHRT (Hybrid Hierarchical RunTime) [Endo, IEEE Cluster 13]
  - Supports transparent data swapping between device memory and host memory
  - Supports oversubscribed execution of processes
  - Works as a wrapper library of MPI and CUDA
    - MPI/CUDA applications run on HHRT with very small modifications

Typical execution w/o HHRT

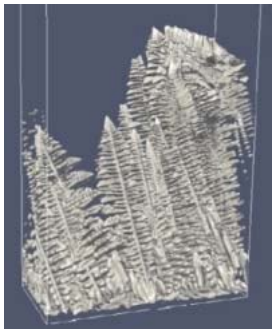


Oversubscribed execution with HHRT

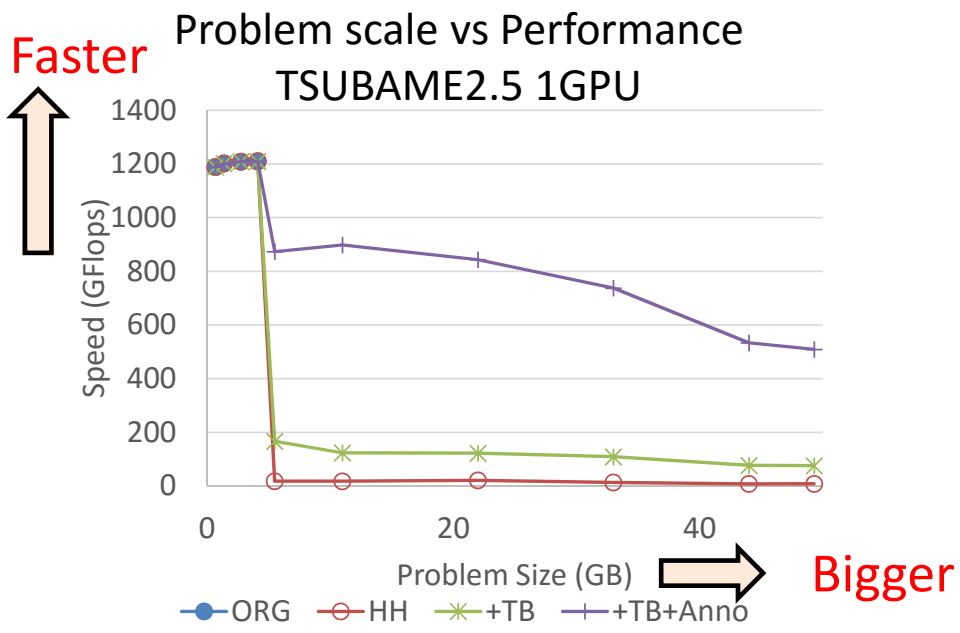


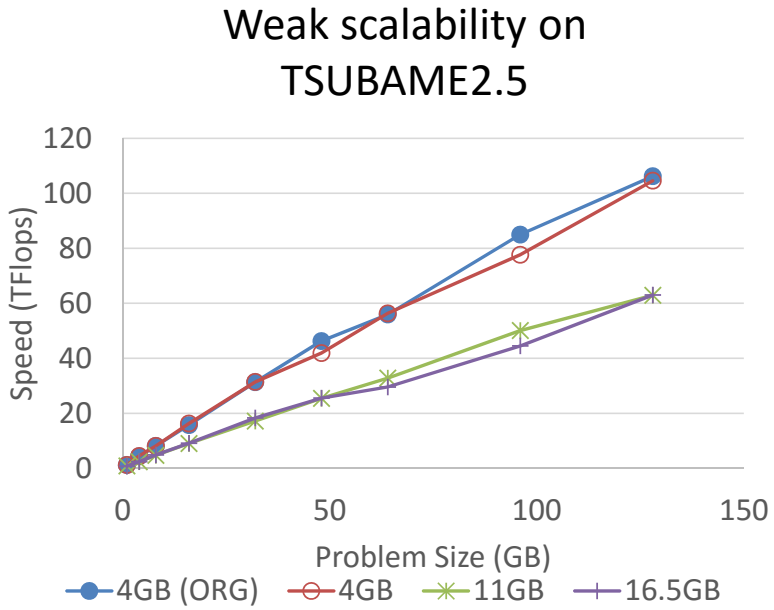
- Several processes time-share the capacity of device memory
- Aggregated data size among local processes can exceed device memory

Performance of Dendrite Simulation Application



- We applied our approach to another stencil application, Dendrite simulation (Gordon Bell Prize 2011)





Applications & Algorithms

Slides by Kengo Nakajima

Information Technology Center
The University of Tokyo

New Frontiers of Computer & Computational Science
towards Post Moore Era
December 22, 2015, Takeda Auditorium, The University of Tokyo

Assumptions & Expectations towards Post-Moore Era

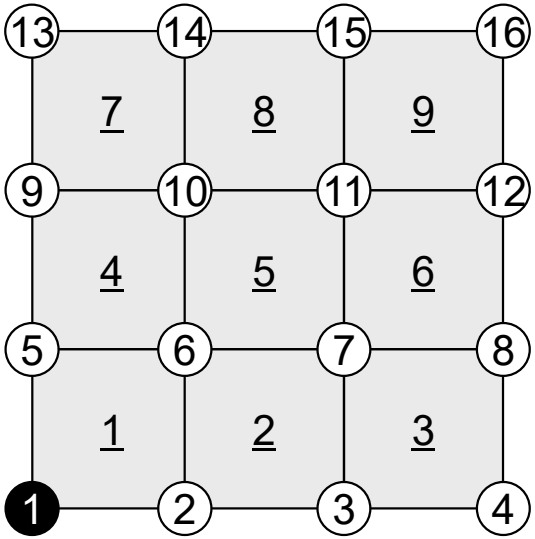
- **Higher Bandwidth**, Larger & Heterogeneous Latency
  - Memory: 3D Stacked Memory
  - Network: Optical Communication
  - Both of Memory & Network will be more hierarchical
- Larger Size of Memory & Cache
- Transaction/Transactional Memory

- Application-Customized Hardware, FPGA
- Large Number of Nodes/Number of Cores per Node
  - under certain constraints (e.g. power, space …)

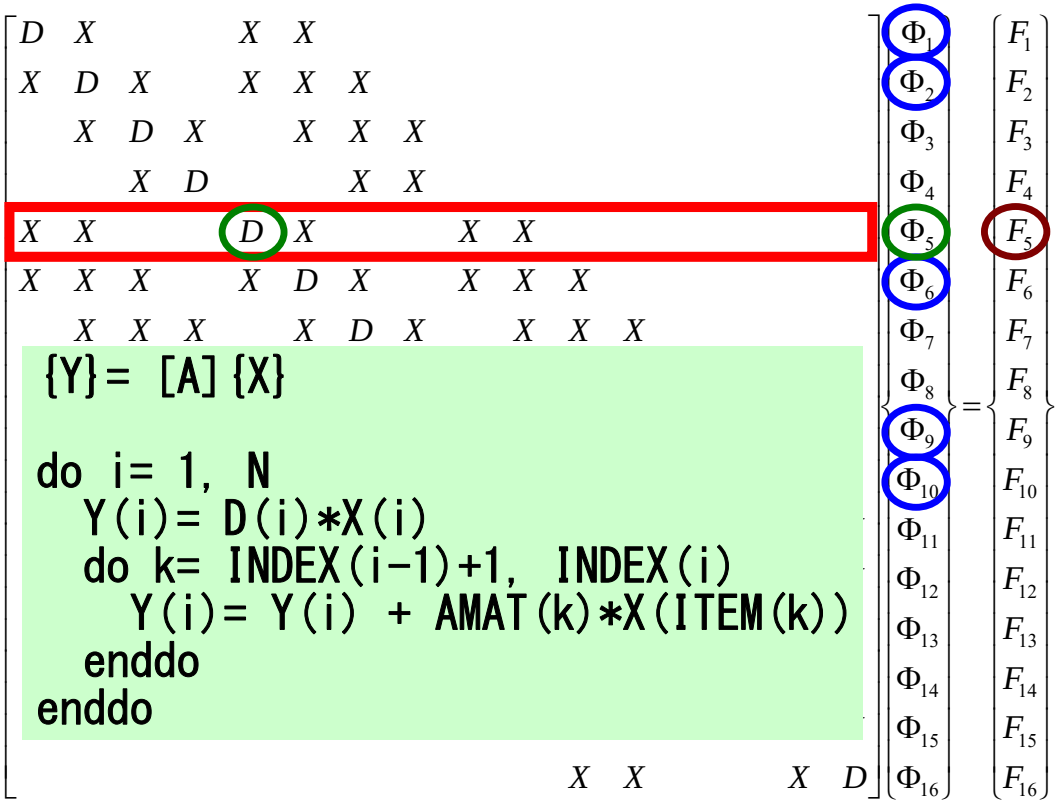
Applications & Algorithms in Post-Moore Era (1/2)

- Compute Intensity ⇒ Data Movement Intensity
  - Non-Blocking Method, Out-of-Core Algorithm
- **Implicit scheme strikes back !**
  - I believe it was never defeated
  - Improvement of performance on sparse matrix computations
  - Big change and advancement are expected in all research areas related to algorithms for sparse matrices including preconditioning
  - **Everything might be easier… but don't relax too much !**

Improvement of performance on sparse matrix computations due to higher memory bandwidth



Sparse Matrices:
- **FEM**
- **Indirect Memory Access**
- **Memory-Bound**



GeoFEM Benchmark: ICCG for FEM Performance of a Node: Flat MPI

	SR11K/J2 Power5+	T2K AMD	FX10	K	Earth Sim 1
Core #/Node	16	16	16	8	8
Peak Performance (GFLOPS)	147.2	147.2	236.5	128.0	64.0
STREAM Triad (GB/s)	101.0	20.0	64.7	43.3	256.0
B/F	0.686	0.136	0.274	0.338	4.00
GeoFEM (GFLOPS)	19.0	4.69	16.0	11.0	25.6
% to Peak	12.9	3.18	6.77	8.59	40.0
LLC/core (MB)	18.0	2.00	0.75	0.75	-

Sparse Solver : Memory-Bound

Applications & Algorithms in Post-Moore Era (2/2)

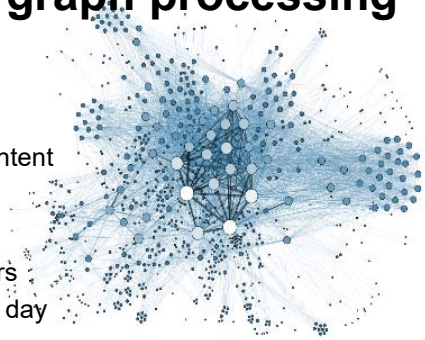
- Hierarchical Methods for Hiding Latency
  - Parallel in Space/Time (PiST)
- Communication/Synchronization Avoiding/Reducing Algorithms
  - Network latency is already a big bottleneck for parallel sparse linear solvers (SpMV)
- Utilization of Manycores
- Power-aware Methods
  - Approximate Computing, Power Management, FPGA

Extreme Big Data Examples

Rates and Volumes are extremely immense

Social NW – large graph processing

- Facebook
  - ~1 billion users
  - Average 130 friends
  - 30 billion pieces of content shared per month
- Twitter
  - 500 million active users
  - 340 million tweets per day
- Internet
  - 300 million new websites per year
  - 48 hours of video to YouTube per minute
  - 30,000 YouTube videos played per second

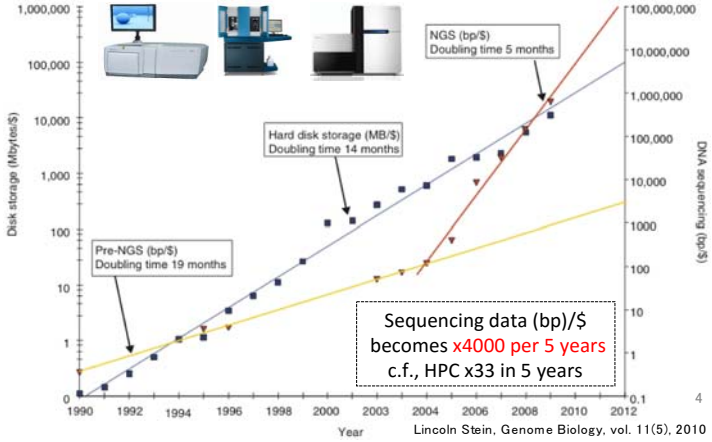


Social Simulation

- Applications
  - Target Area: Planet (Open Street Map)
  - 7 billion people
- Input Data
  - Road Network for Planet: 300GB (XML)
  - Trip data for 7 billion people 10KB (1trip) x 7 billion = 70TB
  - Real-Time Streaming Data (e.g., Social sensor, physical data)
- Simulated Output for 1 Iteration
  - 700TB

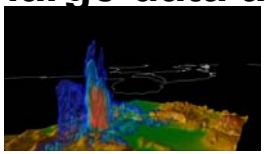
Genomics advanced sequence matching

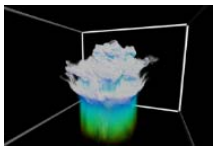
Impact of new generation sequencers

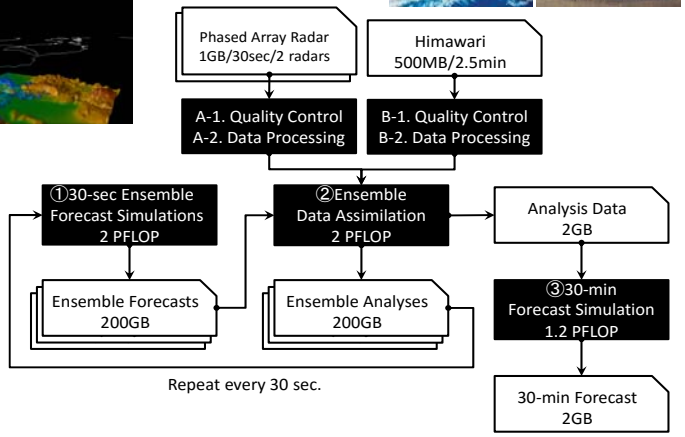


Lincoln Stein, Genome Biology, vol. 11(5), 2010

Weather – real time large data assimilation







NOT simply mining Tbytes Silo Data

Peta~Zetabytes Data

Ultra High-BW Data Stream

Highly Unstructured, Irregular

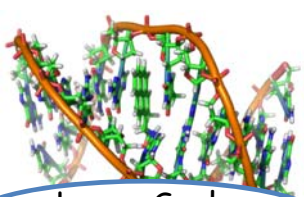
Complex correlations between data from multiple sources

Extreme Capacity, Bandwidth, Compute All Required

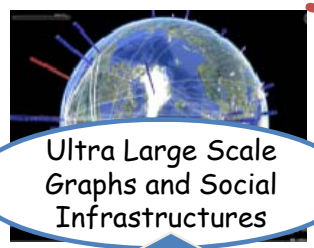
JST-CREST "Extreme Big Data" Project (2013-2018)

Future Non-Silo Extreme Big Data Scientific Apps

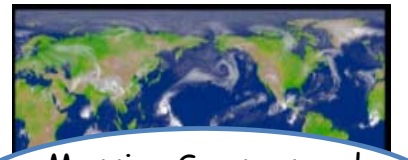
Given a top-class supercomputer, how fast can we accelerate next generation big data c.f. Clouds?



Large Scale Metagenomics



Ultra Large Scale Graphs and Social Infrastructures

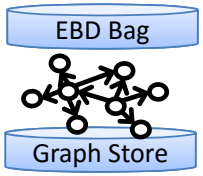


Massive Sensors and Data Assimilation in Weather Prediction

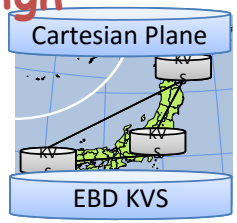
Co-Design

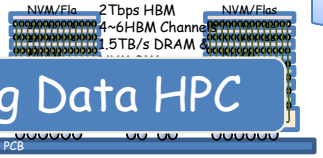
Co-Design

Co-Design



EBD System Software incl. EBD Object System





Exascale Big Data HPC

Convergent Architecture (Phases 1~4)
Large Capacity NVM, High-Bisection NW

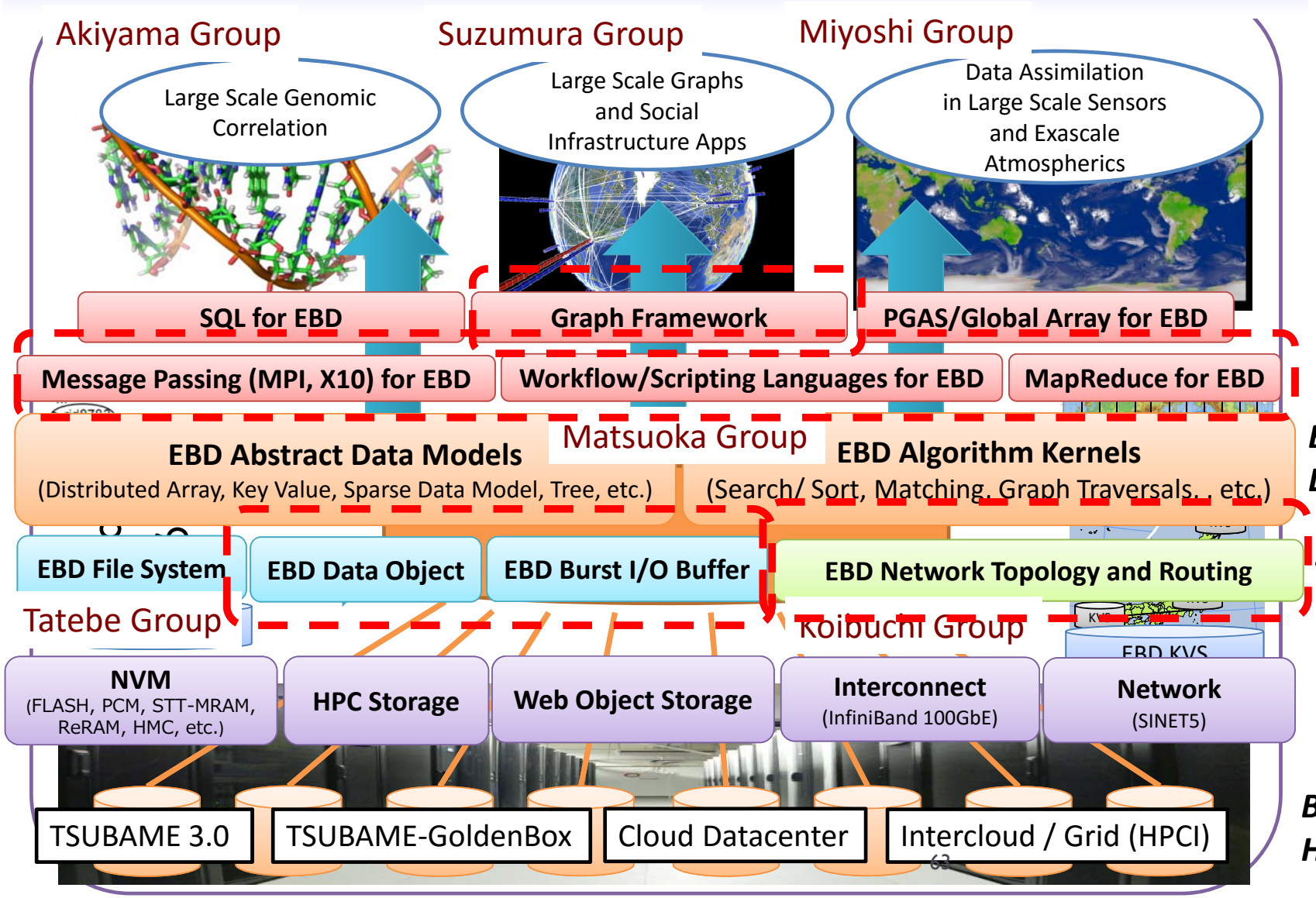
Cloud IDC
Very low BW & Efficiency
Highly available, resilient

Supercomputers
Compute&Batch-Oriented
More fragile

Issues regading Architectural, algorithmic, and system software evolution?

Use of GPUs?

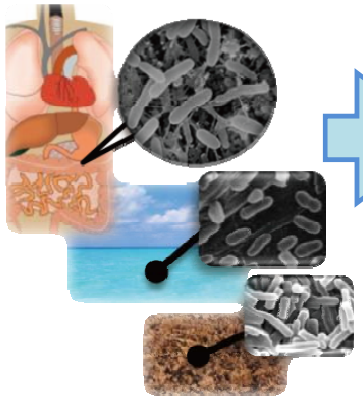
100,000 Times Fold EBD "Convergent" System Architecture



EBD App1: Ultra-fast and High-sensitive Homology Search for Metagenomics [Akiyama Group]

GOAL

1. To develop ultra-fast and high-sensitive metagenomic analyzer.
2. To generalize the invented scheme, as a common tool for "EBD vs. EBD".



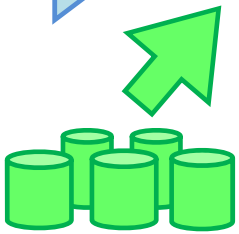
Environmental samples

...CCTTATCTTCG...

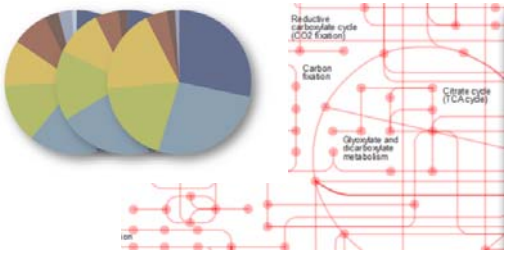
...CCACATAAACT...

...ATGGTCGATGTT...

DNA sequencing



Reference Databases



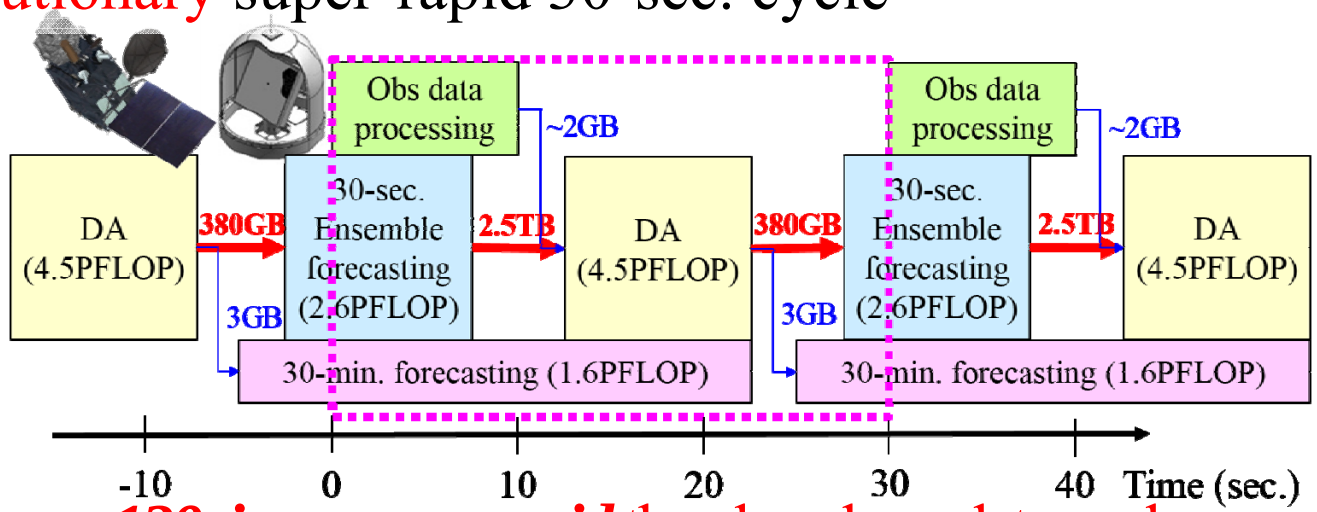
Phylogenetic composition and Metabolic pathway analysis

EBD App2: Miyoshi Group (Weather Forecast Application)

Big Data Assimilation
for severe weather forecast

Goal：Pinpoint (100-m resol.) forecast of severe local weather by updating 30-min forecast every 30 sec!

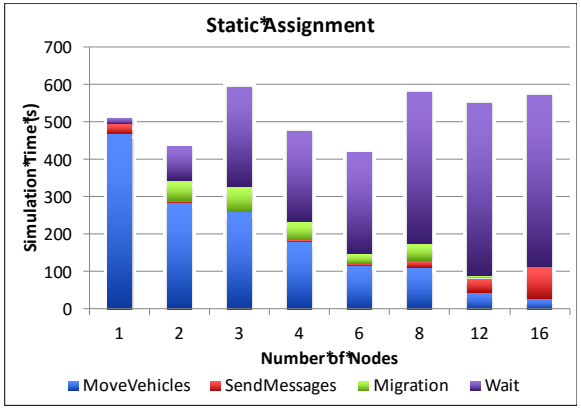
Revolutionary super-rapid 30-sec. cycle



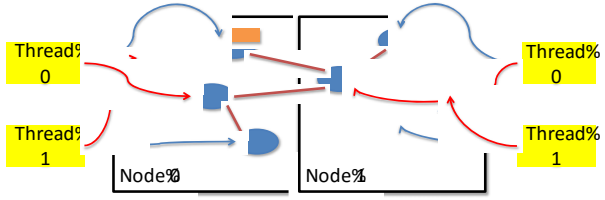
120 times more rapid than hourly update cycles

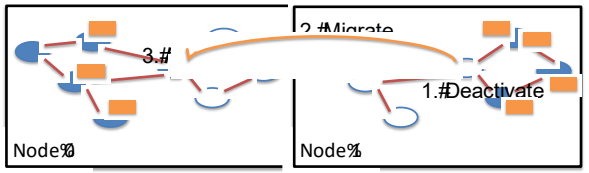
Performance Optimization for Agent-Based Traffic Simulation [Kanezashi et. al., WSC 2015]

- Background
  - Parallel and distributed computation systems are indispensable for large-scale simulations.
  - But there are many performance overheads especially load imbalance.



- Optimization Methods
  1. Inner-node agent assignments
  2. Across-node agent assignments
  3. Synchronization reduction





Acceleration of EBD Processing (1)

- Large Capacity – Multi-Terabytes, Petabytes, Exabytes
- Kernel algorithms for discrete data – graph, sort, etc.
  - EBD Characteristics
    - Sparse and random data structure
    - Involve frequent and abundant data transfer

    } Implies low latency and high bandwidth access
  - EBD Solutions (research)
    - High capacity at low power: non-volatile memory, deep memory hierarchy
    - High bandwidth: fast on-package memory + memory hierarchy+ Supercomputer Network (>100Gbps injection, Petabits bisection) + **bandwidth reducing algorithms for EBD**
    - Low Latency
      - latency reduction => memory 3-D stacking, fast on-package memory + low latency network
      - Latency hiding => many core + many threading + **latency reducing algorithms for EBD**

    } Our research: define & invent EBD architecture + algorithm + system SW

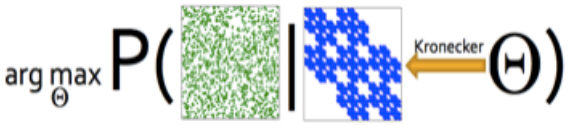
Acceleration of EBD Processing (2)

- Classification algorithms – statistical modeling/optimization, Machine Learning
  - EBD Characteristics: iterative numerical optimization
    - Kernel may be sparse (e.g., SVM) or dense (e.g., Deep Learning)
    - Parallelism difficult due to massive sample size (10~100 billion images)
  - EBD Solutions (our research)
    - Approach: Employ traditional and new HPC/supercomputer parallelization and acceleration strategies
    - Sparse algorithms – high bandwidth processors (e.g., GPU) w/stacked memory and on-package memory + memory hierarchy + supercomputing network + **bandwidth reducing algorithms (sparse linear algebra)**
    - Dense algorithms – many-core high FLOPS processor (e.g., GPU) + **algorithmic advances for strong scaling**
    - High volume data – **utilize "burst buffer" technology (incl. Clouds)**

Limited showing today

Graph500 “Big Data” Benchmark

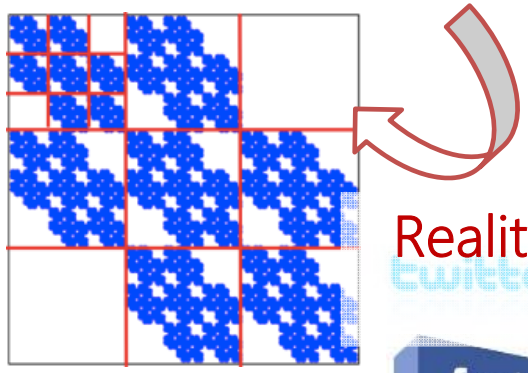
Kronecker graph BSP Problem



A: 0.57, B: 0.19
C: 0.19, D: 0.05

1	1	0
1	1	1
0	1	1

G_1



G_4 adjacency matrix

November 15, 2010

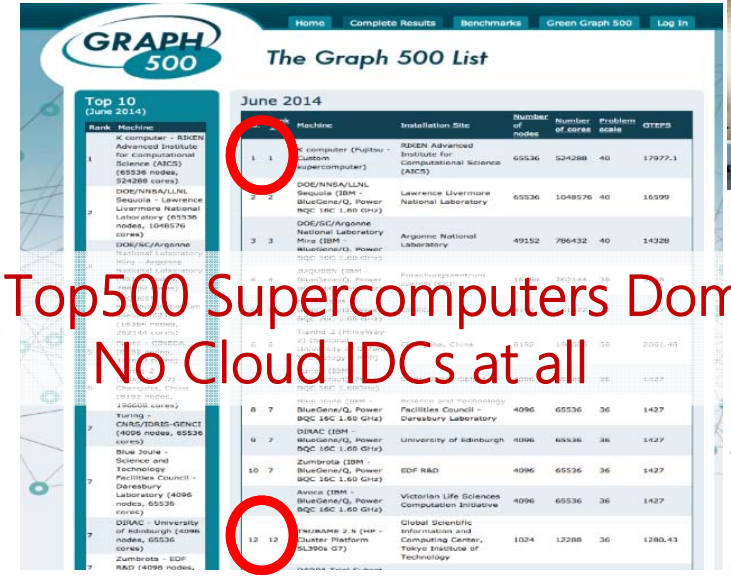
Graph 500 Takes Aim at a New Kind of HPC

Richard Murphy (Sandia NL => Micron)

“ **I expect that this ranking may at times look very different from the TOP500 list. Cloud architectures will almost certainly dominate a major chunk of part of the list.**”

The 8th Graph500 List (June2014): K Computer #1, TSUBAME2 #12

Koji Ueno, Tokyo Institute of Technology/RIKEN AICS

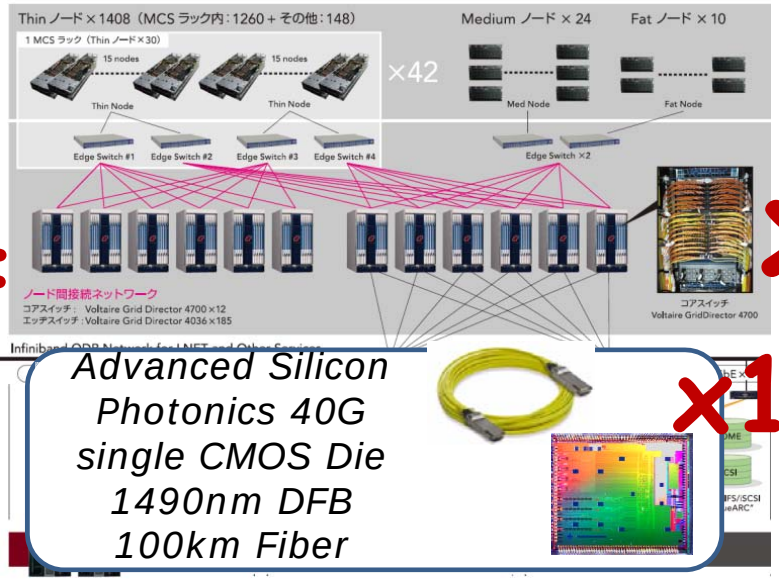


Reality: Top500 Supercomputers Dominate
No Cloud IDCs at all

Supercomputer Tokyo Tech. Tsubame 2.0 #4 Top500 (2010)

A Major Northern Japanese Cloud Datacenter (2013)

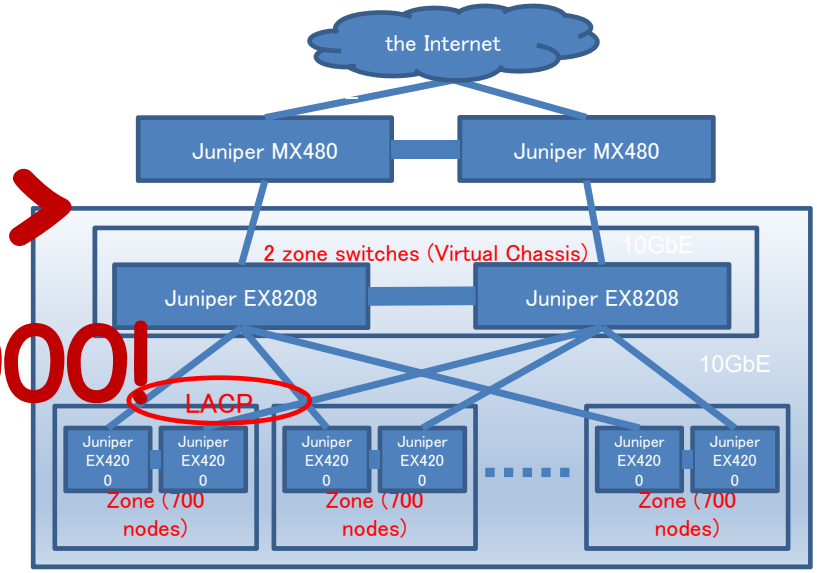
~=



Advanced Silicon Photonics 40G single CMOS Die 1490nm DFB 100km Fiber

>>

x1000!



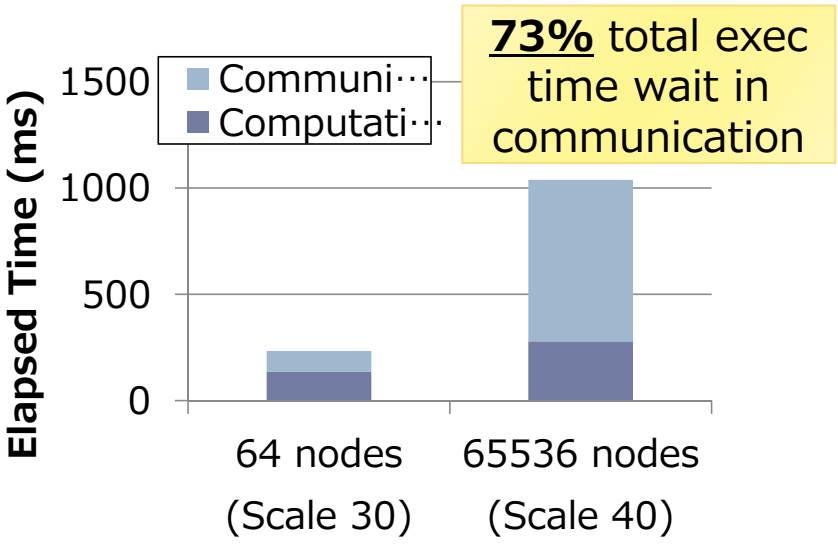
Entire Global Internet Average Data BW ~200 Tbps (source: CISCO)

~1500 nodes compute & storage
Full Bisection Multi-Rail Optical Network
Injection 80GBps/Node
Bisection 220Terabps

8 zones, Total 5600 nodes,
Injection 1GBps/Node
Bisection 160Gigabps

The Graph500 – June 2014 and June 2015

K Computer #1 Tokyo Tech[EBD CREST] Univ. Kyushu [Fujisawa Graph CREST], Riken AICS, Fujitsu



73% total exec time wait in communication

88,000 nodes,
700,000 CPU Cores
1.6 Petabyte mem
20GB/s Tofu NW

K computer

LLNL-IBM Sequoia
1.6 million CPUs
1.6 Petabyte mem

IBM

List	Rank	GTEPS	Implementation
November 2013	4	5524.12	Top-down only
June 2014	1	17977.05	Efficient hybrid
November 2014	2		Efficient hybrid
June 2015	**1**	**38621.4**	Hybrid + Node Compression

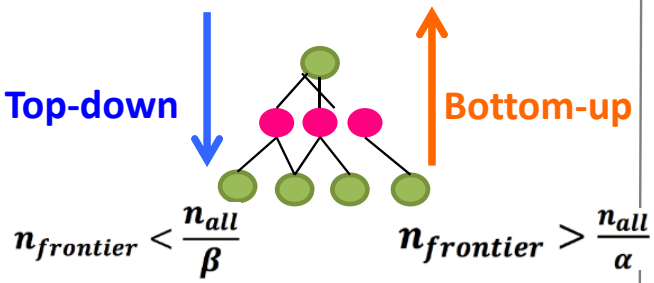
*Problem size is weak scaling "Brain-class" graph

Large Scale Graph Processing Using NVM [Iwabuchi, IEEE BigData2014]

1. Hybrid-BFS (Beamer'11)

EBD Algorithm Kernels

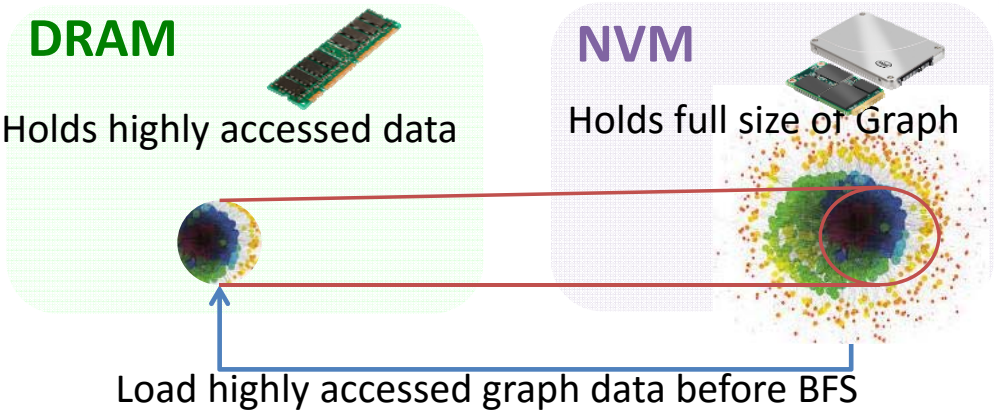
Switching two approaches



$n_{frontier} < \frac{n_{all}}{\beta}$ $n_{frontier} > \frac{n_{all}}{\alpha}$

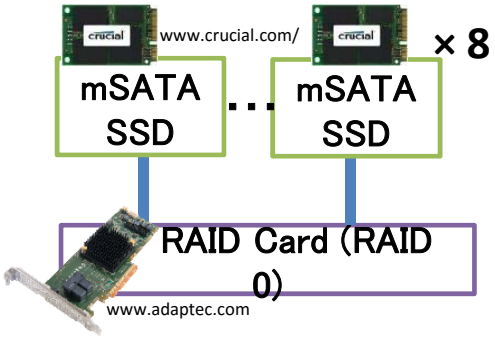
# of frontiers:$n_{frontier}$, # of all vertices:n_{all}, Parameter : α, β

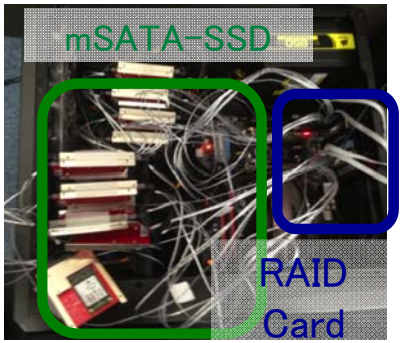
2. Proposal

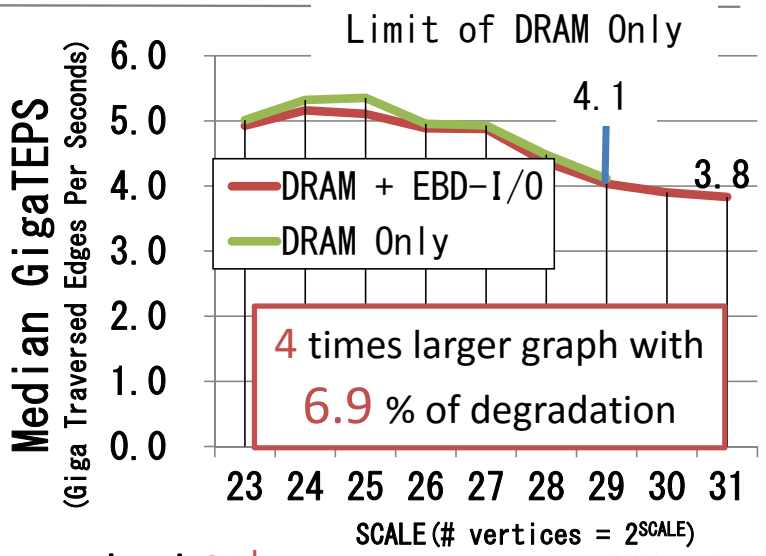


3. Experiment

CPU	Intel Xeon E5-2690 × 2
DRAM	256 GB
NVM	EBD-I/O 2TB × 2







Ranked 3rd

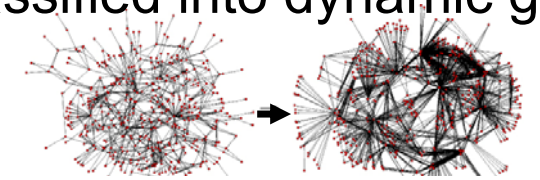
in **Green Graph500 (June 2014)**

Distributed Large-Scale Dynamic Graph Data Store

Keita Iwabuchi[1, 2], Scott Sallinen[3], Roger Pearce[2],
Brian Van Essen[2], Maya Gokhale[2], Satoshi Matsuoka[1]
1. Tokyo Institute of Technology (Tokyo Tech)
2. Lawrence Livermore National Laboratory (LLNL)
3. University of British Columbia

Dynamic Graphs (temporal graph)

- the structure of a graph changes dynamically over time
- many real-world graphs are classified into dynamic graph



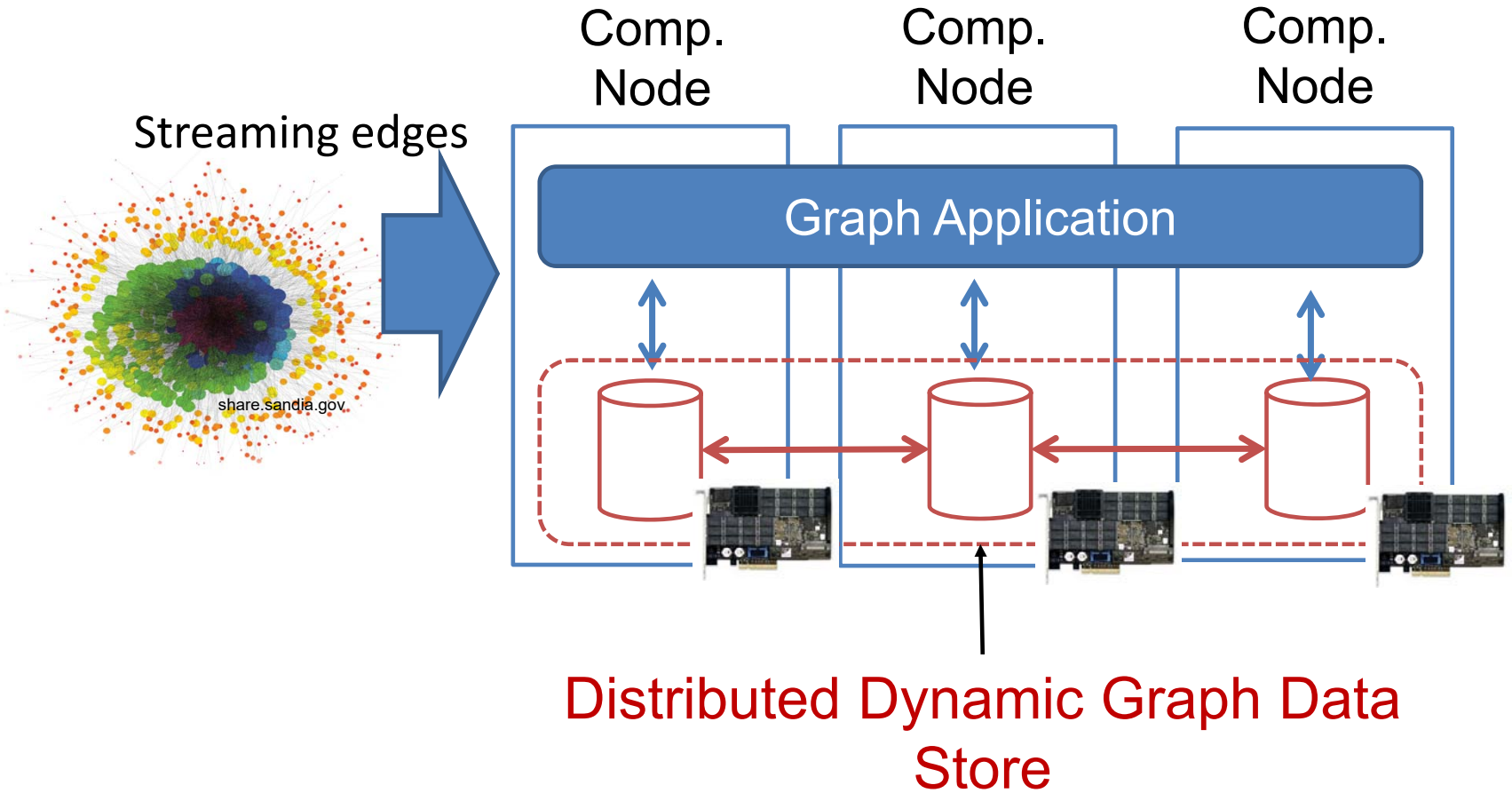
Source: Jakob Enemark and Kim Sneppen, "Gene duplication models for directed networks with limits on growth", Journal of Statistical Mechanics: Theory and Experiment 2007

Sparse Large Scale-free

- social network, genome analysis, WWW, etc.
  - e.g., Facebook manages 1.39 billion active users as of 2014, with more than 400 billion edges

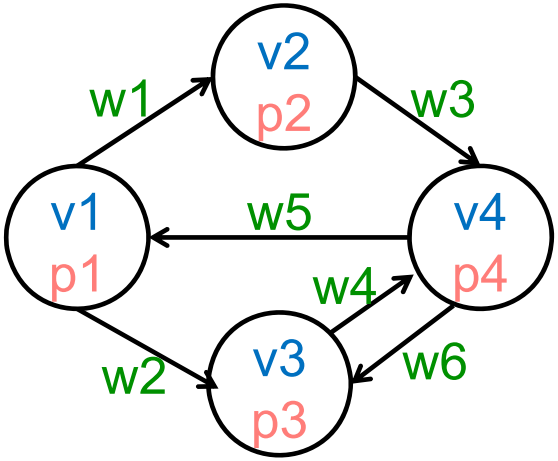
- Most studies for large graphs have not focused on a dynamic graph data structure, but rather a static one, such as Graph 500
- Even with the large memory capacities of HPC systems, many graph applications require additional out-of-core memory (this part is still at an early stage)

Developing a distributed dynamic graph store for data intensive supercomputers equipped with locally attached NVRAM

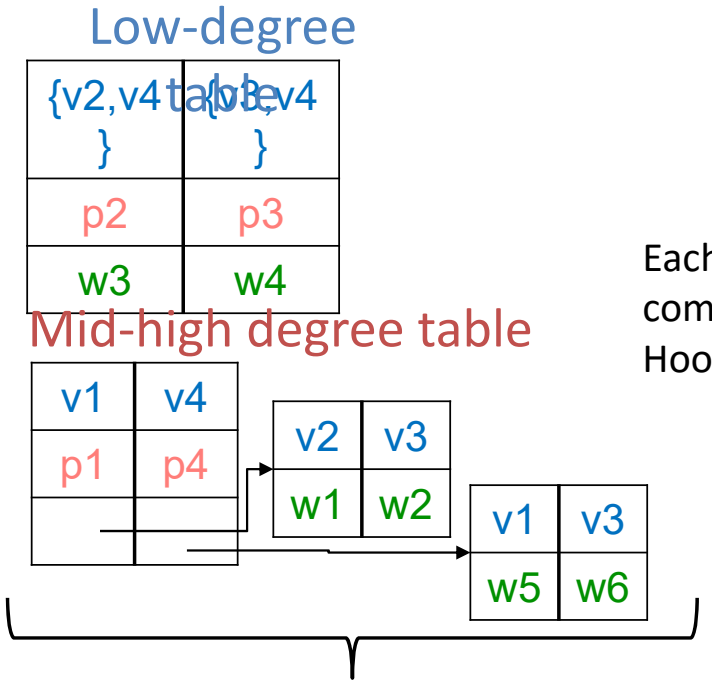


Degree Aware Dynamic Graph Data Store (DegAwareRHH)

- Degree aware data structures, where low-degree vertices are compactly represented
- Use Robin Hood Hashing[1] because of its locality properties to minimize the number of accesses to NVRAM, reducing page misses.



Vertex ID
Vertex property
Edge weight



Each table is composed of Robin Hood Hashing

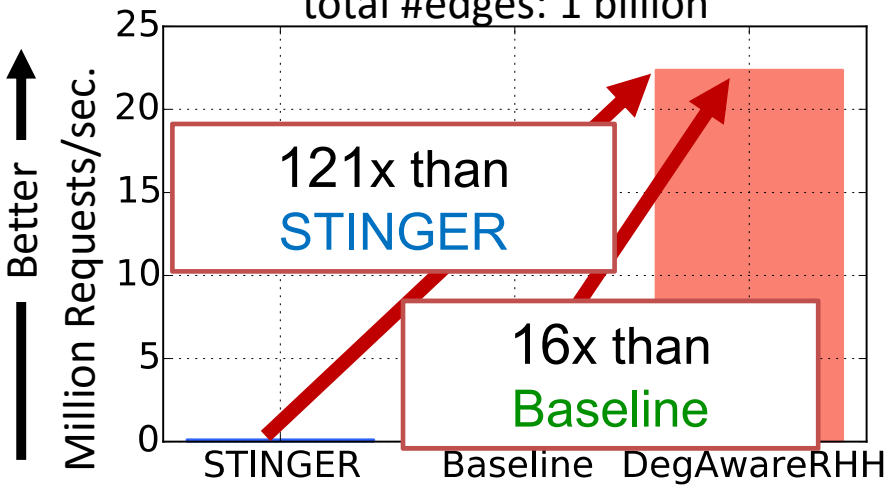
Extend DegAwareRHH for distributed-memory using a async. MPI communication framework[2][3]

[2] R. Pearce, et al, “Scaling techniques for massive scale-free graphs in distributed (external) memory,” IPDPS’ 13

Dynamic Large-Scale Graph Construction (on-memory)

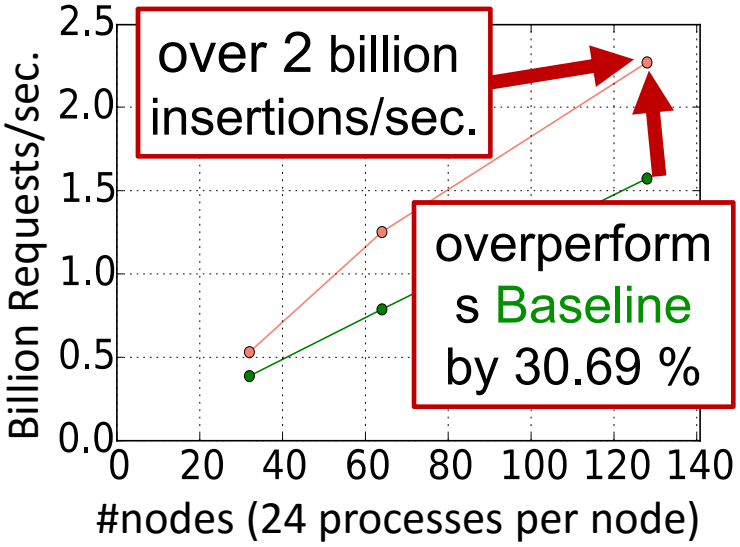
- **STINGER**: a state-of-the-art shared-memory dynamic graph processing framework developing at Georgia Tech
- **Baseline**: a baseline model using *Boost.Interprocess*
- **DegAwareRHH**: our proposed dynamic graph store

Edge insertion and deletion
(single node, 24 threads/processes)
total #edges: 1 billion



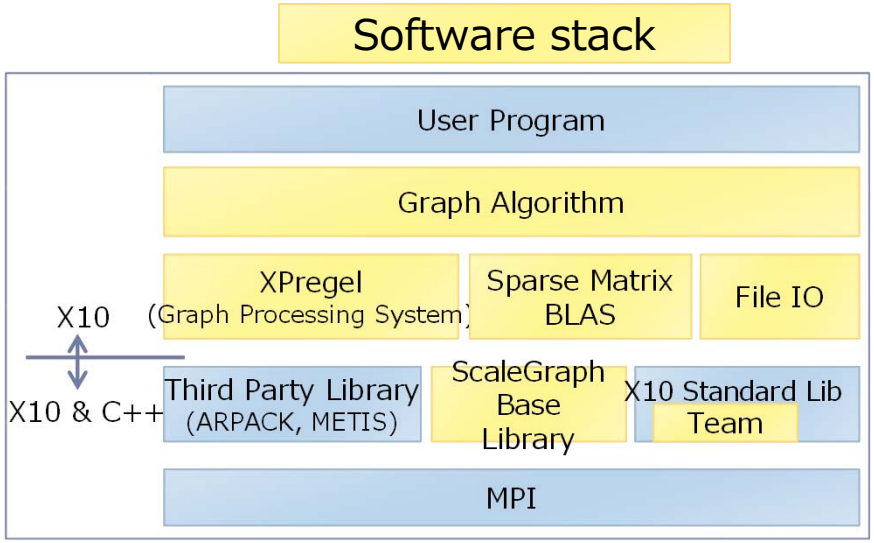
Due to a skewness of the data set (RMAT graph), DegAwareRHH overperforms the both implementations significantly

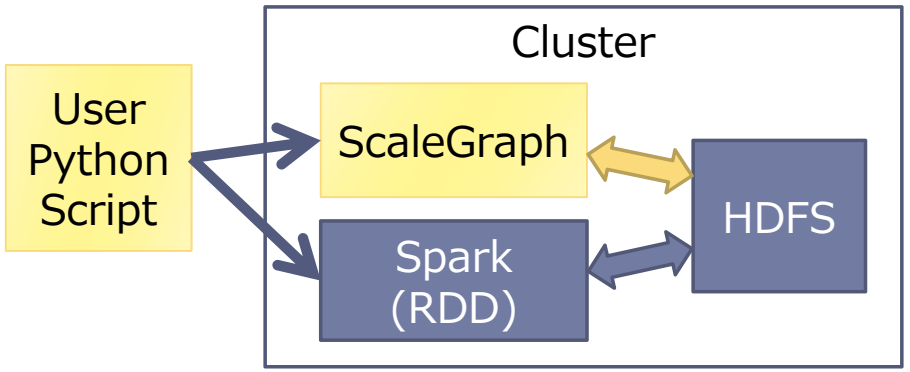
Edge insertion
total #edges: 128 billion



ScaleGraph Large-scale Graph Processing Framework enhanced w/ User-Friendly Python / Spark Interface

- ScaleGraph [Suzumura]
  - X10-based open source **Highly Scalable Large Scale Graph Analytics Library** beyond the scale of billions of vertices and edges on Distributed Systems
    - **XPregel**: Pregel-based bulk synchronous parallel graph processing framework
    - Built-in graph algorithms (Centrality, Connected Component, Clustering, etc.)
- NEW Development: Python Interface
  - Allow users to use ScaleGraph with Spark* by easy python interface





*Apache Spark: http://spark.apache.org/

Estimated Compute Resource Requirements for Deep Learning [Source: Preferred Network Japan Inc.]

To complete the learning phase in one day

P:Peta
E:Exa
F:Flops

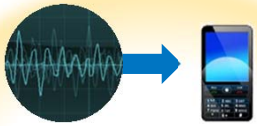
Image/Video Recognition

10P（Image）～ 10E（Video） Flops
学習データ：1億枚の画像 10000クラス分類
数千ノードで6ヶ月 [Google 2015]

Bio / Healthcare

100P ～ 1E Flops
一人あたりゲノム解析で約10M個のSNPs
100万人で100PFlops、1億人で1EFlops

Image Recognition



10P～ Flops
1万人の5000時間分の音声データ
人工的に生成された10万時間の
音声データを基に学習 [Baidu 2015]

Auto Driving

1E～100E Flops
自動運転車１台あたり1日 1TB
10台～1000台, 100日分の走行データの学習

Robots / Drones

1E～100E Flops
1台あたり年間1TB
100万台～1億台から得られた
データで学習する場合

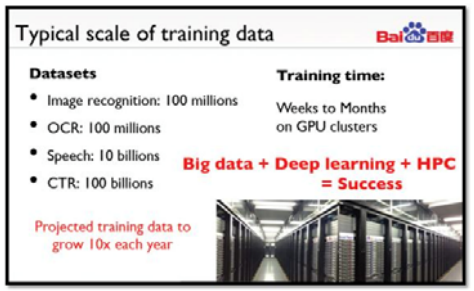
機械学習、深層学習は学習データが大きいほど高精度になる
現在は人が生み出したデータが対象だが、今後は機械が生み出すデータが対象となる

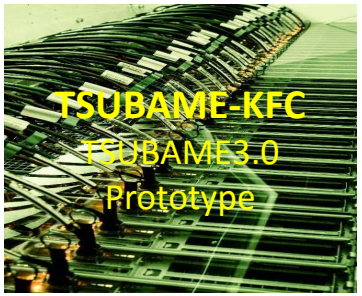
各種推定値は1GBの学習データに対して1日で学習するためには
1TFlops必要だとして計算

10PF (2015) → 100PF (2020) → 1EF (2025) → 10EF → 100EF (2030)

NEW App 2015: EBD Co-Design for Deep Learning Applications

- Deep Learning IS HPC!
  - Training models – mostly dense MatVec
  - Data Access for training target data sets
  - Sharing updated training parameters in neural networks
- Goals
  - Accelerate DL applications in EBD architectures ?
    - Extreme-scale Parallelization, Fast Interconnects, Storage I/O, etc.
  - Performance bottlenecks of multi-node parallel DL algorithms on current HPC systems ?
- Current Status
  - Official Collaboration w/DENSO IT Lab to be signed October
  - Profiling based bottleneck identification and performance optimization of a real DL application on TSUBAME
  - > 100 million images, 1500 GPUs (6 Pflops) 1 week grand challenge run
  - Compete w/Google, MS, Baidu etc. in ILSVRC in ImageNet

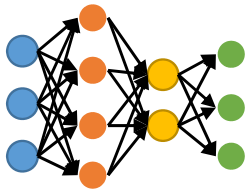




Many companies (ex. Baidu, etc.) employ GPU-based Cluster Architectures, similar to TSUBAME2 & KFC

DENSO

Real DL Applications



I/O | Comm | Calc

Performance Model

Feed Back

GPU-based Distributed Sorting [Shamoto, IEEE BigData 2014, IEEE Trans. Big Data 2015]

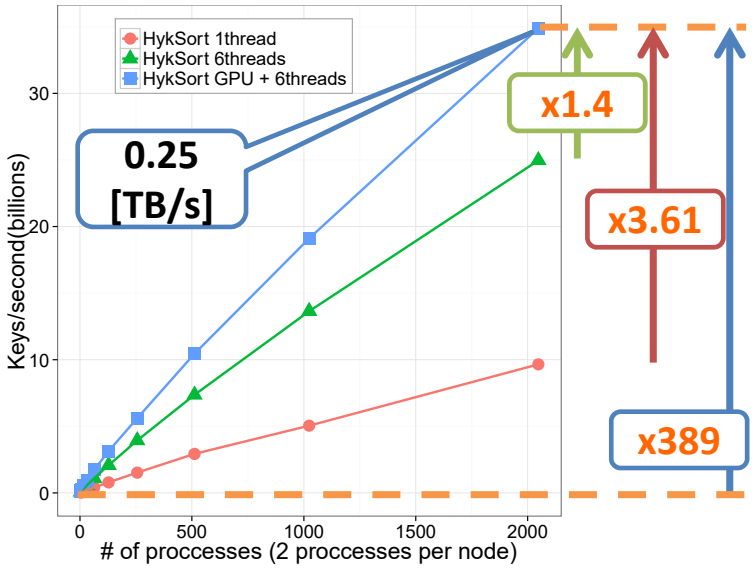
EBD Algorithm Kernels

- Sorting: Kernel algorithm for various EBD processing
- Fast sorting methods
  - Distributed Sorting: Sorting for distributed system
    - Splitter-based parallel sort
    - Radix sort
    - Merge sort
  - Sorting on heterogeneous architectures
    - Many sorting algorithms are accelerated by many cores and high memory bandwidth.
- Sorting for large-scale heterogeneous systems remains unclear

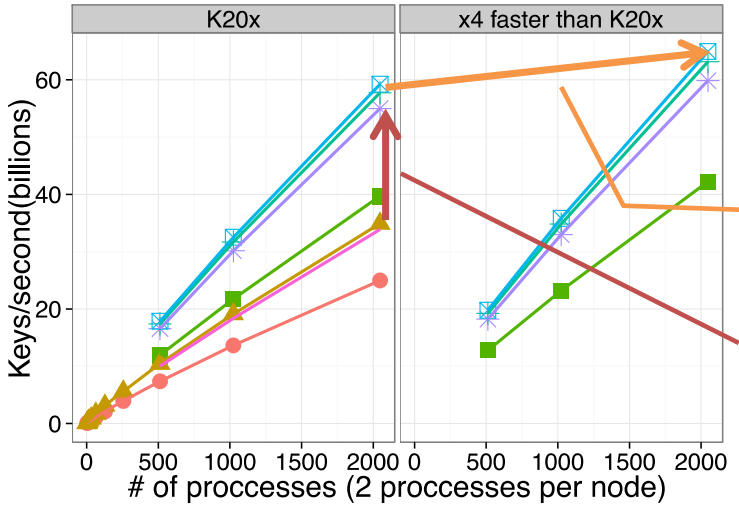
- We develop and evaluate bandwidth and latency reducing GPU-based HykSort on TSUBAME2.5 via latency hiding
  - Now preparing to release the sorting library

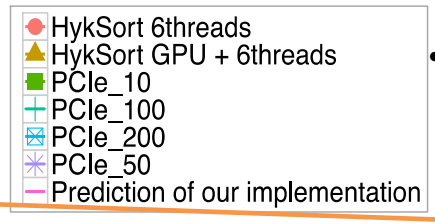
GPU implementation of splitter-based sorting (HykSort)

- Weak scaling performance (Grand Challenge on TSUBAME2.5)
  - 1 ~ 1024 nodes (2 ~ 2048 GPUs)
  - 2 processes per node
  - Each node has 2GB 64bit integer
- Yahoo/Hadoop Terasort: 0.02[TB/s]
  - Including I/O



Performance prediction





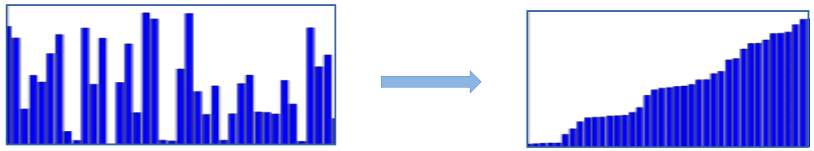
- PCIe_#: #GB/s bandwidth of interconnect between CPU and GPU

x2.2 speedup compared to CPU-based implementation when the # of PCI bandwidth increase to 50GB/s

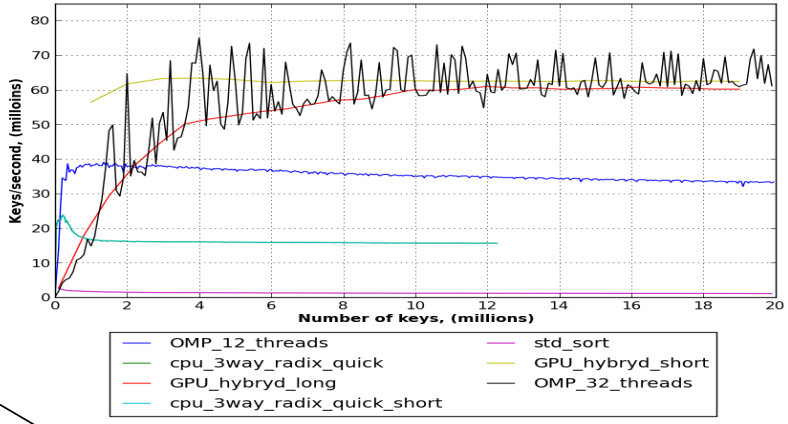
8.8% reduction of overall runtime when the accelerators work 4 times faster than K20x

EBD Algorithm Kernels

Efficient Parallel Sorting Algorithm for Variable-Length Keys



- Comparison-based sorts inefficient for long/variable-length keys (like strings)
- Better way: examining individual characters (based on MSD Radix sort algorithm)
- Hybrid parallelization scheme: combining data-parallel and task-parallel stages



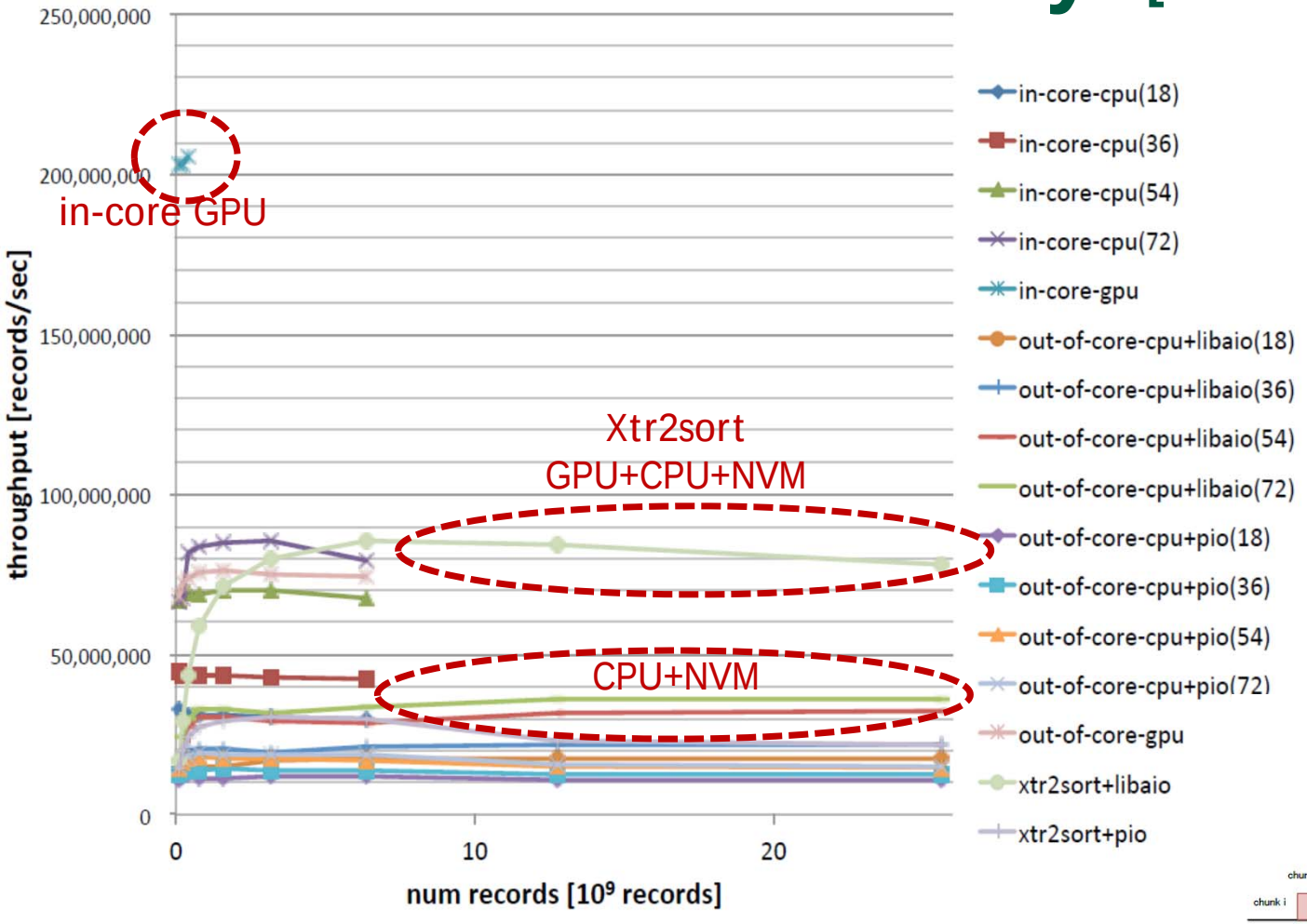
apple
apricot
banana
kiwi

70 M keys/second sorting throughput on 100bytes strings

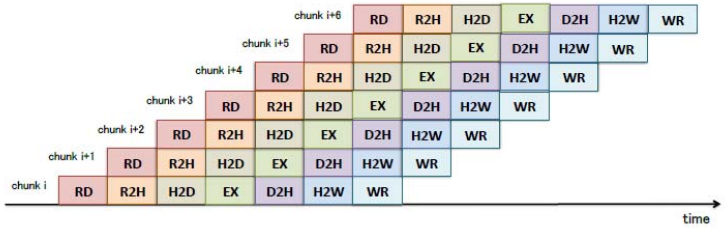
Aleksandr Drozd, Miquel Pericàs, Satoshi Matsuoka. Efficient String Sorting on Multi- and Many-Core Architectures. *in Proceedings of IEEE 3rd International Congress on Big Data. Anchorage, USA, August 2014*

Aleksandr Drozd, Miquel Pericàs, Satoshi Matsuoka. MSD Radix String Sort on GPU: Longer Keys, Shorter Alphabets *in proceedings of 第142回ハイパフォーマンスコンピューティング合同研究発表会 (HOKKE-21)*

GPU + NVM + PCIe SSD Sorting our new Xtr2sort library [H.Sato et.al. SC15 Poster]

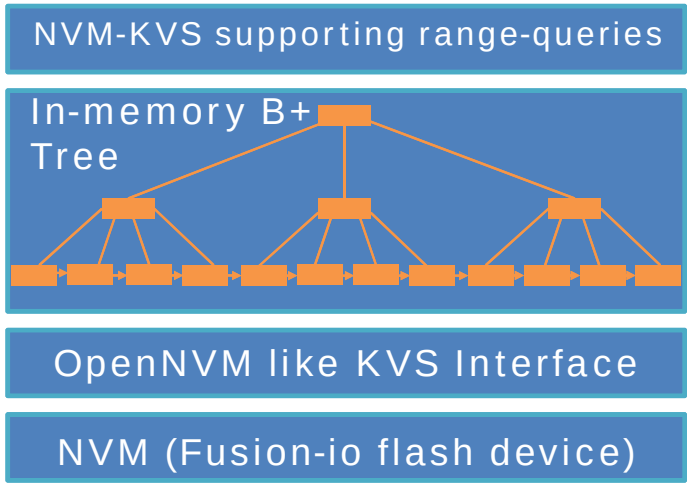


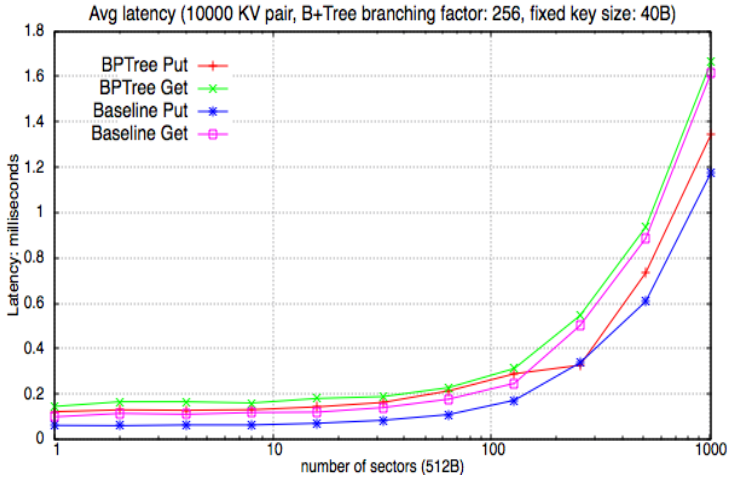
Single Node Xeon
- 2 socket 36 cores
- 128GB DDR4
- K40 GPU (12GB)
- SSD PCIe card (2.4TB)



Concurrent B+Tree Index for Native NVM-KVS [Jabri]

- Enable range-queries support for KVS running natively on NVM like fusionio ioDrive
- Design of Lock-free concurrent B+Tree
  - Lock-free operations – search, insert and delete
    - Dynamic rebalancing of the Tree
    - Nodes to be split or merged are frozen until replaced by new nodes
- Asynchronous interface using future/promise in C++11/14





JOIN US for POST MOORE-EXTREME BIG DATA RESEARCH!

- Whole day of device, hardware, software and algorithms research for post-Moore
- Join our efforts, for innovating the next generation of IT research
- New ideas and research topics welcome!